(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 3 779 881 A1

(12) **EUROPEAN PATENT APPLICATION**
published in accordance with Art. 153(4) EPC

(43) Date of publication:
**17.02.2021 Bulletin 2021/07**

(21) Application number: **19781105.2**

(22) Date of filing: **11.03.2019**

(51) Int Cl.:
***G06T 7/55*** *(2017.01)*

(86) International application number:
**PCT/CN2019/077669**

(87) International publication number:
**WO 2019/192290 (10.10.2019 Gazette 2019/41)**

(84) Designated Contracting States:
**AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR**
Designated Extension States:
**BA ME**
Designated Validation States:
**KH MA MD TN**

(30) Priority: **04.04.2018 CN 201810301988**

(71) Applicant: **TENCENT TECHNOLOGY (SHENZHEN) COMPANY LIMITED**
**Shenzhen, Guangdong 518057 (CN)**

(72) Inventors:
- **JIE, Zequn**
  **Shenzhen, Guangdong 518057 (CN)**
- **LING, Yonggen**
  **Shenzhen, Guangdong 518057 (CN)**
- **LIU, Wei**
  **Shenzhen, Guangdong 518057 (CN)**

(74) Representative: **Eisenführ Speiser**
**Patentanwälte Rechtsanwälte PartGmbB**
**Postfach 31 02 60**
**80102 München (DE)**

# (54) METHOD FOR DETERMINING DEPTH INFORMATION AND RELATED DEVICE

(57) The present application discloses a method for determining depth information, comprising: acquiring a t$^{th}$ left-eye similarity matching from a left-eye image to a right-eye image and a t$^{th}$ right-eye similarity matching from the right-eye image to the left-eye image, wherein t is an integer greater than 1; performing, via a neural network model, processing on the t$^{th}$ left-eye similarity matching and a (t-1)$^{th}$ left-eye attention map so as to acquire a t$^{th}$ left-eye parallax image; performing, via the neural network model, processing on the t$^{th}$ right-eye similarity matching and a (t-1)$^{th}$ right-eye attention map so as to acquire a t$^{th}$ right-eye parallax image; and determining first depth information according to the t$^{th}$ left-eye parallax image, and determining second depth information according to the t$^{th}$ right-eye parallax image. The present application further discloses a device for determining depth information. The present application employs recursive learning to fully consider complementary information of two eyes so as to continuously correct a two-eye parallax image, such that for a region in which two-eye matching is difficult to perform, depth information errors can be effectively reduced.

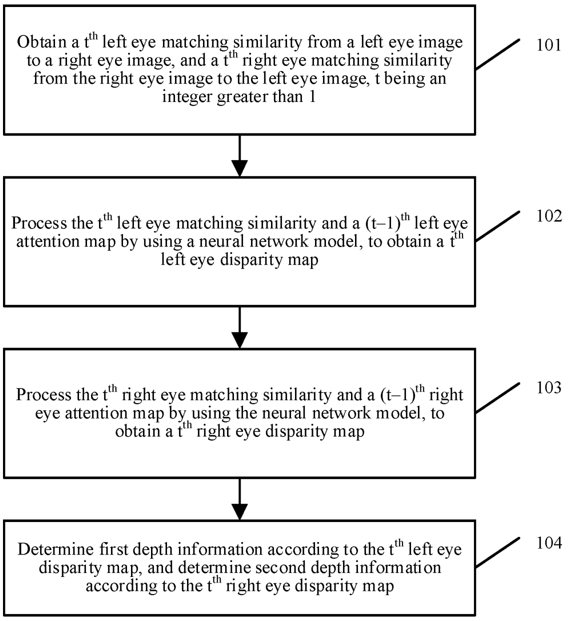
Obtain a t$^{th}$ left eye matching similarity from a left eye image to a right eye image, and a t$^{th}$ right eye matching similarity from the right eye image to the left eye image, t being an integer greater than 1
101
Process the t$^{th}$ left eye matching similarity and a (t–1)$^{th}$ left eye attention map by using a neural network model, to obtain a t$^{th}$ left eye disparity map
102
Process the t$^{th}$ right eye matching similarity and a (t–1)$^{th}$ right eye attention map by using the neural network model, to obtain a t$^{th}$ right eye disparity map
103
Determine first depth information according to the t$^{th}$ left eye disparity map, and determine second depth information according to the t$^{th}$ right eye disparity map
104

FIG. 2

Printed by Jouve, 75001 PARIS (FR)

## Description

RELATED APPLICATION

**[0001]** This application claims priority to Chinese Patent Application No. 201810301988.3, filed with the National Intellectual Property Administration, PRC on April 4, 2018 and entitled "METHOD FOR DETERMINING DEPTH INFORMATION AND RELATED DEVICE ", which is incorporated herein by reference in its entirety.

FIELD OF THE TECHNOLOGY

**[0002]** This application relates to the field of computer processing, and in particular, to determining of depth information.

BACKGROUND OF THE DISCLOSURE

**[0003]** A disparity is a difference between directions in which a viewer views the same object at two different locations. For example, after reaching out a finger in front of eyes, if the person first closes the right eye and views the finger with the left eye, and then closes the left eye and views the finger with the right eye, the person finds that a location of the finger relative to a background object is changed, which is a disparity in viewing the same point from different angles.

**[0004]** Currently, in a process of predicting depth information of an object, matching similarities from the left eye to the right eye at different disparities need to be first predicted, and then disparity prediction is performed on a left eye image by using the matching similarities from the left eye to the right eye at the different disparities. In this way, the depth information of the object is determined.

**[0005]** However, for a region (e.g., a repetitive region, a texture-less region, and an edge of a complex object) that is difficult for matching between the two eyes, the depth information is prone to a relatively large error if only the matching similarities from the left eye to the right eye at the different disparities are used.

SUMMARY

**[0006]** Embodiments of this application provide a depth information determining method and a related apparatus, which, through recursive learning, may continuously correct disparity maps of two eyes by fully using complementary information of the two eyes, to effectively reduce an error of depth information for a region difficult to match between the two eyes.

**[0007]** A first aspect of the embodiments of this application provides a depth information determining method, including:

obtaining a t$^{th}$ left eye matching similarity from a left eye image to a right eye image, and a t$^{th}$ right eye matching similarity from the right eye image to the left eye image, t being an integer greater than 1;

processing the t$^{th}$ left eye matching similarity and a (t-1)$^{th}$ left eye attention map by using a neural network model, to obtain a t$^{th}$ left eye disparity map;

processing the t$^{th}$ right eye matching similarity and a (t-1)$^{th}$ right eye attention map by using the neural network model, to obtain a t$^{th}$ right eye disparity map; and

determining first depth information according to the t$^{th}$ left eye disparity map, and determining second depth information according to the t$^{th}$ right eye disparity map.

**[0008]** A second aspect of the embodiments of this application provides a depth information determining apparatus, including:

an obtaining module, configured to obtain a t$^{th}$ left eye matching similarity from a left eye image to a right eye image, and a t$^{th}$ right eye matching similarity from the right eye image to the left eye image, t being an integer greater than 1;

a processing module, configured to process the t$^{th}$ left eye matching similarity obtained by the obtaining module and a (t-1)$^{th}$ left eye attention map by using a neural network model, to obtain a t$^{th}$ left eye disparity map,

the processing module being further configured to process the t$^{th}$ right eye matching similarity obtained by the obtaining module and a (t-1)$^{th}$ right eye attention map by using the neural network model, to obtain a t$^{th}$ right eye disparity map; and a determining module, configured to determine first depth information according to the t<sup>th</sup> left eye disparity map obtained by the processing module through processing, and determine second depth information according to the t<sup>th</sup> right eye disparity map obtained by the processing module through processing.

**[0009]** A third aspect of the embodiments of this application provides a depth information determining apparatus, including a memory, a processor, and a bus system,
the memory being configured to store a program;
the processor being configured to execute the program in the memory, to perform the following operations:

obtaining a t<sup>th</sup> left eye matching similarity from a left eye image to a right eye image, and a t<sup>th</sup> right eye matching similarity from the right eye image to the left eye image, t being an integer greater than 1;

processing the t<sup>th</sup> left eye matching similarity and a (t-1)<sup>th</sup> left eye attention map by using a neural network model, to obtain a t<sup>th</sup> left eye disparity map;

processing the t<sup>th</sup> right eye matching similarity and a (t-1)<sup>th</sup> right eye attention map by using the neural network model, to obtain a t<sup>th</sup> right eye disparity map; and

determining first depth information according to the t<sup>th</sup> left eye disparity map, and determining second depth information according to the t<sup>th</sup> right eye disparity map; and

the bus system being configured to connect the memory and the processor for communication between the memory and the processor.

**[0010]** A fourth aspect of the embodiments of this application provides a computer-readable storage medium, storing instructions, the instructions, when running on a computer, causing the computer to perform the methods according to the foregoing aspects.

**[0011]** It may be seen from the foregoing technical solutions that the embodiments of this application have the following advantages:

**[0012]** In the embodiments of this application, a depth information determining method is provided, including: obtaining a t<sup>th</sup> left eye matching similarity from a left eye image to a right eye image, and a t<sup>th</sup> right eye matching similarity from the right eye image to the left eye image; then processing the t<sup>th</sup> left eye matching similarity and a (t-1)<sup>th</sup> left eye attention map by using a neural network model, to obtain a t<sup>th</sup> left eye disparity map, and processing the t<sup>th</sup> right eye matching similarity and a (t-1)<sup>th</sup> right eye attention map by using the neural network model, to obtain a t<sup>th</sup> right eye disparity map; and finally determining first depth information according to the t<sup>th</sup> left eye disparity map, and determining second depth information according to the t<sup>th</sup> right eye disparity map. In this way, disparity maps of two eyes may be obtained by using a neural network model and attention maps of the two eyes obtained through previous learning, and current attention maps of the two eyes obtained through learning according to the currently obtained disparity maps of the two eyes are used for next disparity maps of the two eyes. Through such recursive learning, the disparity maps of the two eyes may be continuously corrected by fully using complementary information of the two eyes, which effectively reduces an error of depth information for a region difficult to match between the two eyes.

## BRIEF DESCRIPTION OF THE DRAWINGS

**[0013]**

FIG. 1A shows schematic disparity maps of two eyes based on recursive learning according to an embodiment of this application.

FIG. 1B is a schematic architectural diagram of a depth information determining apparatus according to an embodiment of this application.

FIG. 2 is a schematic diagram of an embodiment of a depth information determining method according to an embodiment of this application.

FIG. 3 is a schematic diagram of comparisons between original maps and model-based depth prediction maps according to an embodiment of this application.

FIG. 4 is a schematic diagram of generating attention maps of two eyes according to an embodiment of this application.

FIG. 5 is a schematic diagram of a recursive binocular disparity network according to an embodiment of this application.

FIG. 6 is a schematic diagram of a convolutional long short-term memory network according to an embodiment of this application.

FIG. 7 is a schematic diagram of an embodiment of a depth information determining apparatus according to an embodiment of this application.

FIG. 8 is a schematic diagram of another embodiment of a depth information determining apparatus according to an embodiment of this application.

FIG. 9 is a schematic structural diagram of a depth information determining apparatus according to an embodiment of this application.

DESCRIPTION OF EMBODIMENTS

**[0014]** Embodiments of this application provide a depth information determining method and a related apparatus, which, through recursive learning, may continuously correct disparity maps of two eyes by fully using complementary information of the two eyes, to effectively reduce an error of depth information for a region difficult to match between the two eyes.

**[0015]** The terms "first", "second", "third", "fourth", and the like (if exists) in the specification and the claims of this application and the foregoing accompanying drawings are used for distinguishing similar objects, and are not necessarily used to describe a particular sequence or order. The data termed in such a way is interchangeable in proper circumstances so that the embodiments of this application described herein can, for example, be implemented in other orders that are different from the order illustrated or described herein. Moreover, the terms "include", "have" and any other variants are intended to cover the non-exclusive inclusion, for example, a process, method, system, product, or device that includes a list of steps or units is not necessarily limited to those expressly listed steps or units, but may include other steps or units not expressly listed or inherent to such a process, method, product, or device.

**[0016]** It is to be understood that, this application may be applied to a facility (e.g., a binocular robot and an unmanned vehicle) equipped with a binocular camera for object depth estimation. In this application, depth values are obtained by obtaining disparities of visual images of two eyes through a deep neural network and then dividing a product of a spacing of the binocular camera and a focal length of the binocular camera by the predicted disparities. Specifically, similarities of matching from an image of one eye to an image of the other eye (e.g., from the left eye to the right eye, and from the right eye to the left eye) at different disparities are first predicted by using a convolutional neural network, to obtain matching similarities at different disparities, and then a cycle of "disparity prediction of the two eyes to comparison between disparity maps of the two eyes" is recursively performed by using a convolutional long short-term memory (ConvLSTM) network. In this cycle, through continuous comparison between disparity maps of the two eyes, complementary information of the left and right visions can be fully used to automatically detect a region (e.g., a repetitive region, a texture-less region, or an edge of a complex object) that is difficult to match in left and right visions, so as to correct and update predicted disparity values of the two eyes, and continuously improve accuracy of disparity prediction, that is, accuracy of a depth.

**[0017]** For left and right visual images photographed by the binocular camera, matching similarities from the left eye image to the right eye image and from the right eye image to the left eye image at different disparities are first predicted by using the convolutional neural network, and then recursive prediction is performed on disparities of the two eyes based on the predicted matching similarities by using the ConvLSTM network. The whole flowchart is shown in FIG. 1A. FIG. 1A shows schematic disparity maps of two eyes based on recursive learning according to an embodiment of this application. It is assumed that both left and right eye images photographed by the binocular camera have a resolution of H*W (H is height, and W is width). As shown in the figure, pixel-level feature extraction is first performed on the left and right eye images by using the convolutional neural network, to obtain H*W*C (C is feature dimension) feature maps respectively for the two images. Then, features in the two H*W*C feature maps are combined based on different disparities in a horizontal direction, to obtain a maximum of $D_{max}$ feature maps (H*W*2C*$D_{max}$ dimensions) at the different disparities. Then, another convolutional neural network whose convolution kernel is 1*1 is used to predict matching similarities of all pixels at the different disparities, and one matching similarity value is obtained based on a 2C input feature. Similarity values of H*W pixels at all $D_{max}$ possible disparities are written in a tensor form, so that one matching similarity of H*W*$D_{max}$ can be predicted from the left eye image to the right eye image and from the right eye image to the left eye image.

**[0018]** Based on the foregoing predicted matching similarity tensors of the two eyes, the ConvLSTM network is used to perform recursive prediction on disparities of the two eyes, so as to obtain a left eye disparity map and a right eye disparity map.

**[0019]** With the development in the past few decades, a stereoscopic vision is applied more widely to fields such as robot vision, aerial mapping, reverse engineering, military application, medical imaging, and industrial detection. FIG. 1B is a schematic architectural diagram of a depth information determining apparatus according to an embodiment of this application. As shown in the figure, the depth information determining apparatus provided in this application may be deployed on a server, and the server transmits a processing result to a target device. Alternatively, the depth information determining apparatus may be directly deployed on the target device. The target device includes, but is not limited to, an (unmanned) automobile, a robot, an (unmanned) airplane, and an intelligent terminal. All the target devices have a binocular stereoscopic vision, which can obtain two images of a to-be-detected object from different locations based on a disparity principle by using an imaging device, and obtain three-dimensional geometric information of the object by calculating a location deviation between corresponding points in the images. The binocular stereoscopic vision combines images obtained by two eyes and observes a difference between the images, so that an obvious depth perception may be obtained, and a correspondence between features may be established, to make the same spatial physical point correspond to mapping points in different images. The difference is referred to as a disparity image.

**[0020]** A disparity of two eyes, sometimes also referred to as a stereoscopic disparity, is a depth cue. A closer distance between an object and an observer causes a larger difference between object images obtained by two eyes. Therefore, the disparity of two eyes is formed. The brain may estimate a distance between the object and the eyes by measuring the disparity.

**[0021]** The following describes a depth information determining method in this application. Referring to FIG. 2, an embodiment of a depth information determining method according to an embodiment of this application includes the following steps 101-104.

**[0022]** 101. Obtain a $t^{th}$ left eye matching similarity from a left eye image to a right eye image, and a $t^{th}$ right eye matching similarity from the right eye image to the left eye image, t being an integer greater than 1.

**[0023]** In this embodiment, a depth information determining apparatus first obtains a left eye image and a right eye image by using a binocular camera, and then calculates a $t^{th}$ left eye matching similarity from the left eye image to the right eye image, and a $t^{th}$ right eye matching similarity from the right eye image to the left eye image, t being an integer greater than 1, which may be considered as a matching similarity obtained for the $t^{th}$ time. The following describes several algorithms for calculating the matching similarity. An actual application includes, but is not limited to, the following listed algorithms.

**[0024]** The first algorithm is a mean absolute difference (MAD) algorithm. The algorithm has a simple idea and relatively high matching precision, and is widely applied to image matching. In a search map S, (i, j) may be taken as the upper left corner, and a submap of M*N is obtained. A similarity between the submap and a template is calculated. The whole search map S is traversed, and in all submaps that can be obtained, a submap most similar to the template map is found and taken as a final matching result.

**[0025]** The second algorithm is a sum of absolute difference (SAD) algorithm. An idea of the SAD algorithm is nearly the same as that of the MAD algorithm, but a similarity measurement formula of the SAD algorithm is slightly different. Details are not described herein.

**[0026]** The third algorithm is a sum of squared difference (SSD) algorithm, also referred to as a difference quadratic sum algorithm. The SSD algorithm is the same as the SAD algorithm, but a similarity measurement formula of the SSD algorithm is slightly different. Details are not described herein.

**[0027]** The fourth algorithm is a normalized cross correlation (NCC) algorithm. Similar to the foregoing algorithms, gradations of a submap and a template map are used to calculate a matching degree between the two maps through a normalized correlation measurement formula.

**[0028]** The fifth algorithm is a sequential similarity detection algorithm (SSDA), which is an improvement of a conventional template matching algorithm, and is tens to hundreds of times faster than the MAD algorithm.

**[0029]** 102. Process the $t^{th}$ left eye matching similarity and a $(t-1)^{th}$ left eye attention map by using a neural network model, to obtain a $t^{th}$ left eye disparity map.

**[0030]** In this embodiment, the depth information determining apparatus inputs the currently ($t^{th}$) obtained left eye matching similarity and a previously ($(t-1)^{th}$) generated left eye attention into a neural network model. The neural network model is generally obtained through training in advance, and the neural network model outputs a current ($t^{th}$) left eye disparity map.

**[0031]** 103. Process the $t^{th}$ right eye matching similarity and a $(t-1)^{th}$ right eye attention map by using the neural network model, to obtain a $t^{th}$ right eye disparity map.

**[0032]** In this embodiment, similarly, the depth information determining apparatus inputs the currently ($t^{th}$) obtained right eye matching similarity and a previously ($(t-1)^{th}$) generated right eye attention map into the neural network model. The neural network model is generally obtained through training in advance, and the neural network model outputs a current ($t^{th}$) right eye disparity map.

**[0033]** It may be understood that step 102 may be performed before, after, or simultaneously with step 103. This is not limited herein.

**[0034]** 104. Determine first depth information according to the $t^{th}$ left eye disparity map, and determine second depth information according to the $t^{th}$ right eye disparity map.

**[0035]** In this embodiment, the depth information determining apparatus determines, according to the $t^{th}$ left eye disparity map outputted by the neural network model, depth information of the $t^{th}$ left eye disparity map (that is, first depth information). Similarly, the depth information determining apparatus determines, according to the $t^{th}$ right eye disparity map outputted by the neural network model, depth information of the $t^{th}$ right eye disparity map (that is, second depth information).

**[0036]** For ease of description, reference is made to FIG. 3 which is a schematic diagram of comparison between original maps and model-based depth prediction maps according to an embodiment of this application. As shown in the figure, a high-quality depth map may be obtained through prediction by using a neural network model provided in this application. In this application, accuracy of binocular object depth estimation can be improved, which plays a decisive role in self driving and work of a facility such as a robot and an unmanned vehicle that are equipped with a binocular camera, and has potential economic benefits.

**[0037]** In the embodiments of this application, a depth information determining method is provided, including: obtaining a $t^{th}$ left eye matching similarity from a left eye image to a right eye image, and a $t^{th}$ right eye matching similarity from the right eye image to the left eye image; then processing the $t^{th}$ left eye matching similarity and a $(t-1)^{th}$ left eye attention map by using a neural network model, to obtain a $t^{th}$ left eye disparity map, and processing the $t^{th}$ right eye matching similarity and a $(t-1)^{th}$ right eye attention map by using the neural network model, to obtain a $t^{th}$ right eye disparity map; and finally determining first depth information according to the $t^{th}$ left eye disparity map, and determining second depth information according to the $t^{th}$ right eye disparity map. In this way, disparity maps of two eyes may be obtained by using a neural network model and attention maps of the two eyes obtained through previous learning, and current attention maps of the two eyes obtained through learning according to the currently obtained disparity maps of the two eyes are used for next disparity maps of the two eyes. Through such recursive learning, the disparity maps of the two eyes may be continuously corrected by fully using complementary information of the two eyes, which effectively reduces an error of depth information for a region difficult to match between the two eyes.

**[0038]** Optionally, based on the embodiment corresponding to FIG. 2, a first optional embodiment of the depth information determining method provided in the embodiments of this application may further include the following steps:

mapping the $t^{th}$ right eye disparity map to a left eye coordinate system, to obtain a $t^{th}$ left eye mapping disparity map;

generating a $t^{th}$ left eye attention map according to the $t^{th}$ left eye mapping disparity map and the $t^{th}$ left eye disparity map;

mapping the $t^{th}$ left eye disparity map to a right eye coordinate system, to obtain a $t^{th}$ right eye mapping disparity map; and

generating a $t^{th}$ right eye attention map according to the $t^{th}$ right eye mapping disparity map and the $t^{th}$ right eye disparity map.

**[0039]** In this embodiment, the depth information determining apparatus generates an attention map by using a mapping disparity map and a disparity map. Specifically, reference is made to FIG. 4 which is a schematic diagram of generating attention maps of two eyes according to an embodiment of this application. As shown in the figure, after a $t^{th}$ right eye disparity map and a $t^{th}$ left eye disparity map are generated by using a neural network model, the $t^{th}$ right eye disparity map may be mapped to a left eye coordinate system, to obtain a $t^{th}$ left eye mapping disparity map, and the $t^{th}$ left eye disparity map is mapped to a right eye coordinate system, to obtain a $t^{th}$ right eye mapping disparity map. The mapping is converting two disparity maps into coordinates of opposite disparity maps. Next, the original $t^{th}$ left eye disparity map and the $t^{th}$ left eye mapping disparity map obtained through conversion are connected, and inputted into a model formed by several simple convolutional layers and transformation layers, to obtain a $t^{th}$ left eye attention map. Similarly, the original $t^{th}$ right eye disparity map and the $t^{th}$ right eye mapping disparity map obtained through conversion are connected, and inputted into the model formed by several simple convolutional layers and transformation layers, to obtain a $t^{th}$ right eye attention map.

**[0040]** An attention map reflects confidences of disparity prediction in different regions after left and right images are compared with each other. A low confidence means that a predicted disparity value of a pixel is not confident. The low-confidence pixel regions that are automatically detected after comparison between left and right eye disparities are usually regions difficult to match between the left and right eyes, such as a repetitive region, a texture-less region, and an edge of a complex object. Therefore, an attention map obtained through t<sup>th</sup> recursive learning can be used for (t+1)th recursive disparity prediction, and a network can purposely correct and update a disparity value of a pixel in a low-confidence region automatically detected by the tth recursion according to this. That is, the attention map may be used for a focus region of a guidance model in a next step.

**[0041]** In addition, in this embodiment of this application, the depth information determining apparatus maps the tth right eye disparity map to a left eye coordinate system, to obtain a tth left eye mapping disparity map, and generates a tth left eye attention map according to the tth left eye mapping disparity map and the tth left eye disparity map. Similarly, a tth right eye attention map may also be obtained. In this way, an attention map obtained through current recursive learning can be used for next recursive disparity prediction, and a network can purposely correct and update a disparity value of a pixel in a low-confidence region automatically detected by the current recursion according to this, thereby improving reliability of the attention maps of the two eyes.

**[0042]** Optionally, based on the first embodiment corresponding to FIG. 2, in a second optional embodiment of the depth information determining method provided in the embodiments of this application, after the determining first depth information according to the tth left eye disparity map, and determining second depth information according to the tth right eye disparity map, the method may further include the following steps:

obtaining a (t+1)th left eye matching similarity from the left eye image to the right eye image, and a (t+1)th right eye matching similarity from the right eye image to the left eye image;

processing the (t+1)th left eye matching similarity and the tth left eye attention map by using the neural network model, to obtain a (t+1)th left eye disparity map;

processing the (t+1)th right eye matching similarity and the tth right eye attention map by using the neural network model, to obtain a (t+1)th right eye disparity map; and

determining third depth information according to the (t+1)th left eye disparity map, and determining fourth depth information according to the (t+1)th right eye disparity map.

**[0043]** In this embodiment, a manner of predicting next depth information is described. Reference is made to FIG. 5 which is a schematic diagram of a recursive binocular disparity network according to an embodiment of this application. As shown in the figure, the recursive binocular disparity network, which may also be referred to as a left-right comparative recurrent (LRCR) model, includes two parallel neural network models. A left-side neural network model generates a tth left eye disparity map by using $X'_t$. $X'_t$ represents a connection result between a tth left eye matching similarity and a (t-1)th left eye attention map. Similarly, a right-side neural network model generates a tth right eye disparity map by using $X''_t$. $X''_t$ represents a connection result between a tth right eye matching similarity and a (t-1)th right eye attention map. Next, a tth left eye attention map and a tth right eye attention map may be predicted by using the tth left eye disparity map and the tth right eye disparity map.

**[0044]** Then, a next cycle may be performed, that is, the left-side neural network model generates a (t+1)th left eye disparity map by using $X'_{t+1}$. $X'_{t+1}$ represents a connection result between a (t+1)th left eye matching similarity and a tth left eye attention map. Similarly, the right-side neural network model generates a (t+1)th right eye disparity map by using $X''_{t+1}$. $X''_{t+1}$ represents a connection result between a (t+1)th right eye matching similarity and a tth right eye attention map. Next, a (t+1)th left eye attention map and a (t+1)th right eye attention map may be predicted by using the (t+1)th left eye disparity map and the (t+1)th right eye disparity map. The rest is deduced by analogy, and details are not described herein.

**[0045]** Further, in this embodiment of this application, after obtaining current depth information of the two eyes, the depth information determining apparatus may further continue to obtain next depth information of the two eyes. In this way, for comparison between the left and right eyes, a convolutional layer and an aggregation layer may be added to the neural network model, so as to generate attention maps of the two eyes. The attention maps of the two eyes are used as an input of a next step, and the LRCR model is started. Left-right mismatching regions may be focused more in the next step, thereby improving prediction accuracy.

**[0046]** Optionally, based on FIG. 2 and the first or second embodiment corresponding to FIG. 2, in a third optional embodiment of the depth information determining method provided in the embodiments of this application, the processing the tth left eye matching similarity and a (t-1)th left eye attention map by using a neural network model, to obtain a tth left eye disparity map may include:

obtaining a tth left eye hidden variable through calculation according to the tth left eye matching similarity and the (t-1)th left eye attention map by using a ConvLSTM network;

obtaining a tth left eye disparity cost according to the tth left eye hidden variable; and

calculating a tth predicted left eye disparity value according to the tth left eye disparity cost, the tth predicted left eye disparity value being used for generating the tth left eye disparity map; and

the processing the tth right eye matching similarity and a (t-1)th right eye attention map by using the neural network model, to obtain a tth right eye disparity map includes:

obtaining a tth right eye hidden variable through calculation according to the tth right eye matching similarity and the (t-1)th right eye attention map by using the ConvLSTM network;

obtaining a tth right eye disparity cost according to the tth right eye hidden variable; and

calculating a tth predicted right eye disparity value according to the tth right eye disparity cost, the tth predicted right eye disparity value being used for generating the tth right eye disparity map.

**[0047]** In this embodiment, in a process of obtaining the tth left eye disparity map, the tth left eye matching similarity and the (t-1)th left eye attention map need to be first inputted into the ConvLSTM network, to obtain the tth left eye hidden variable through calculation. Then the tth left eye disparity cost is obtained according to the tth left eye hidden variable. Finally, the tth predicted left eye disparity value is calculated according to the tth left eye disparity cost. Obtaining the tth predicted left eye disparity value means that the tth left eye disparity map may be generated. Similarly, a manner of generating the tth right eye disparity map is similar to the manner of generating the tth left eye disparity map. Details are not described herein.

**[0048]** For ease of understanding, reference is made to FIG. 6 which is a schematic diagram of a ConvLSTM network according to an embodiment of this application. As shown in the figure, each black line transmits a complete vector, which is outputted from one node and inputted to another node. Circles represent pointwise operations, such as summation of vectors. A matrix is a neural network layer obtained through learning. Combined lines represent connections of vectors, and branched lines represent that content is duplicated, and then is distributed to different locations. If there is only the upper horizontal line, information cannot be added or deleted, and information addition or deletion is implemented by using a structure called gates. The gates may selectively allow information to pass through, which is mainly implemented by using a neural layer of sigmoid and an operation of pointwise multiplication. Each element of an output (which is a vector) of the neural layer of sigmoid is a real number from 0 to 1, which represents a weight (or proportion) for allowing corresponding information to pass through. For example, 0 represents "disallowing any information to pass through", and 1 represents "allowing all information to pass through". A tanh layer represents a repetitive structural module.

**[0049]** The ConvLSTM network protects and controls information by using the structure shown in FIG. 6. The three gates are respectively an input gate, a forget gate, and an output gate.

**[0050]** Further, in this embodiment of this application, the ConvLSTM network is used to process the tth left eye matching similarity and the (t-1)th left eye attention map, to obtain the tth left eye disparity map, and process the tth right eye matching similarity and the (t-1)th right eye attention map, to obtain the tth right eye disparity map. In this way, based on predicted matching similarities of the two eyes, recursive prediction is performed on the disparity maps of the two eyes by using the ConvLSTM network. The ConvLSTM network has strong capabilities of sequence modeling and information processing as a conventional recursive neural network, and also can effectively extract information in neighborhood of each pixel space, so as to integrate spatial context information.

**[0051]** Optionally, based on the third embodiment corresponding to FIG. 2, in a fourth optional embodiment of the depth information determining method provided in the embodiments of this application, the obtaining a tth left eye hidden variable through calculation according to the tth left eye matching similarity and the (t-1)th left eye attention map by using a ConvLSTM network may include:

calculating the tth left eye hidden variable in the following manner:

$$i'_t=\sigma(W_{xi}*X'_t+W_{hi}*H'_{t-1}+W_{ci}\circ C'_{t-1}+b_i);$$

$$f'_t=\sigma(W_{xf}*X'_t+W_{hf}*H'_{t-1}+W_{cf}\circ C'_{t-1}+b_f);$$

$$o'_t=\sigma(W_{xo}*X'_t+W_{ho}*H'_{t-1}+W_{co}\circ C'_{t-1}+b_o);$$

$$C'_t=f'_t\circ C'_{t-1}+i'_t\circ \tanh(W_{xc}*X'_t+W_{hc}*H'_{t-1}+b_c);$$

and

$$H'_t=o'_t\circ \tanh(C'_t),$$

where $i'_t$ represents a network input gate of a t$^{th}$ left eye recursion, * represents multiplication of vectors, ° represents a convolution operation, $\sigma$ represents a sigmoid function, $W_{xi}$, $W_{hi}$, $W_{ci}$, and $b_i$ represent model parameters of the network input gate, $X'_t$ represents the t$^{th}$ left eye matching similarity and the (t-1)$^{th}$ left eye attention map, $f'_t$ represents a forget gate of the t$^{th}$ left eye recursion, $W_{xf}$, $W_{hf}$, $W_{cf}$, and $b_f$ represent model parameters of the forget gate, o't represents an output gate of the t$^{th}$ left eye recursion, $W_{xo}$, $W_{ho}$, $W_{co}$, and $b_o$ represent model parameters of the output gate, $C'_t$ represents a memory cell of the t$^{th}$ left eye recursion, $C'_{t-1}$ represents a memory cell of a (t-1)$^{th}$ left eye recursion, tanh represents a hyperbolic tangent, $H'_{t-1}$ represents a (t-1)$^{th}$ left eye hidden variable, and $H'_t$ represents the t$^{th}$ left eye hidden variable; and

the obtaining a t$^{th}$ right eye hidden variable through calculation according to the t$^{th}$ right eye matching similarity and the (t-1)$^{th}$ right eye attention map by using the ConvLSTM network may include:

calculating the t$^{th}$ right eye hidden variable in the following manner:

$$i''_t=\sigma(W_{xi}*X''_t+W_{hi}*H''_{t-1}+W_{ci}\circ C''_{t-1}+b_i);$$

$$f''_t=\sigma(W_{xf}*X''_t+W_{hf}*H''_{t-1}+W_{cf}\circ C''_{t-1}+b_f);$$

$$o''_t=\sigma(W_{xo}*X''_t+W_{ho}*H''_{t-1}+W_{co}\circ C''_{t-1}+b_o);$$

$$C''_t=f'_t\circ C''_{t-1}+i'_t\circ \tanh(W_{xc}*X''_t+W_{hc}*H''_{t-1}+b_c);$$

and

$$H''_t=o''_t\circ \tanh(C''_t),$$

where

$i''_t$ represents a network input gate of a t$^{th}$ right eye recursion, $X''_t$ represents the t$^{th}$ right eye matching similarity and the (t-1)$^{th}$ right eye attention map, $f'_t$ represents a forget gate of the t$^{th}$ right eye recursion, $o'_t$ represents an output gate of the t$^{th}$ right eye recursion, $C''_t$ represents a memory cell of the t$^{th}$ right eye recursion, $C''_{t-1}$ represents a memory cell of a (t-1)$^{th}$ right eye recursion, $H''_{t-1}$ represents a (t-1)$^{th}$ right eye hidden variable, and $H''_t$ represents the t$^{th}$ right eye hidden variable.

**[0052]** In this embodiment, calculation of hidden variables of the two eyes are specifically described with reference to formulas, and the ConvLSTM network obtains information by using an input gate, a forget gate, and an output gate.

**[0053]** The first step in the ConvLSTM network is to decide information that is to be discarded. This decision is completed by using a forget gate. The gate reads $H'_{t-1}$ (or $H''_{t-1}$) and $X'_t$ (or $X''_t$), and outputs a value from 0 to 1 to each number in a cell state $C'_{t-1}$ (or $C''_{t-1}$). 1 represents "all retained", and 0 represents "all discarded". $H'_{t-1}$ (or $H''_{t-1}$) represents an output of a previous cell, $X'_t$ (or $X''_t$) represents an input of a current cell, and $\sigma$ represents a sigmod function.

**[0054]** The next step is to decide a quantity of pieces of new information added to the cell state. Two substeps are needed to implement this step. First, a sigmoid layer called "input gate layer" decides information that needs to be updated, and a tanh layer generates a vector, that is, candidate content for update. Next, the two portions are combined, to update the cell state. $C'_{t-1}$ (or $C''_{t-1}$) is updated to C't (or C"t). The old state is multiplied by $f'_t$ (or $f''_t$), and information that is determined to be discarded is discarded.

**[0055]** Finally, a value that is to be outputted needs to be determined. This output is based on the cell state, and is a version after filtering. First, a sigmoid layer is run to determine which portion of the cell state is to be outputted. Then, the cell state is processed by using tanh (to obtain a value from -1 to 1), and the value is multiplied by an output of the sigmoid gate. Finally, only the portion that is determined to be outputted is outputted.

**[0056]** Further, in this embodiment of this application, a specific manner of calculating the t^th left eye hidden variable and the t^th right eye hidden variable is described. The hidden variables of the two eyes may be obtained by using a calculation relationship provided by the ConvLSTM network. In this way, reliability of hidden variable calculation can be effectively improved, and an operability basis for implementing the solution is provided.

**[0057]** Optionally, based on the third embodiment corresponding to FIG. 2, in a fifth optional embodiment of the depth information determining method provided in the embodiments of this application, the obtaining a t^th left eye disparity cost according to the t^th left eye hidden variable may include:

processing the t^th left eye hidden variable by using at least two fully connected layers, to obtain the t^th left eye disparity cost; and

the obtaining a t^th right eye disparity cost according to the t^th right eye hidden variable may include:
processing the t^th right eye hidden variable by using at least two fully connected layers, to obtain the t^th right eye disparity cost.

**[0058]** In this embodiment, the t^th left eye hidden variable may be inputted into at least two fully connected layers, and the at least two fully connected layers outputs the t^th left eye disparity cost. Similarly, the t^th right eye hidden variable is inputted into at least two fully connected layers, and the at least two fully connected layers outputs the t^th right eye disparity cost.

**[0059]** Specifically, each node of a fully connected layer is connected to all nodes of a previous layer, to combine features that are previously extracted. Due to a full-connection characteristic of the fully connected layer, the fully connected layer generally has the most parameters. The fully connected layer indeed has many parameters. In a forward calculation process, that is, a linear process of weighted summation, each output of the fully connected layer may be obtained by multiplying each node of a previous layer by a weight coefficient W and then adding a bias value b. Assuming that there are 50*4*4 input neuron nodes and 500 output nodes, 50*4*4*500=400000 weight coefficients W and 500 bias parameters b are needed in total.

**[0060]** A connected layer is actually a convolution operation in which a size of a convolution kernel is a size of previous layer features. A result obtained after the convolution is a node, and the node corresponds to a point in the fully connected layer. Assuming that an output size of the last convolutional layer is 7*7*512, a size of a fully connected layer connected to the convolutional layer is 1*1*4096. A connected layer is actually a convolution operation in which a size of a convolution kernel is a size of previous layer features. A result obtained after the convolution is a node, and the node corresponds to a point in the fully connected layer. If the fully connected layer is converted into a convolutional layer, there are 4096 filters in total. Each filter includes 512 convolution kernels. A size of each convolution kernel is 7*7, and an output size is 1*1*4096. If a fully connected layer of 1*1*4096 is added behind, parameters of a converted convolutional layer corresponding to the fully connected layer include: 4096 filters, each filter includes 4096 convolution kernels, a size of each convolution kernel is 1*1, and an output size is 1*1*4096. That is, features are combined to perform calculation of 4096 classification scores, and an obtained correct category has the highest score.

**[0061]** Further, in this embodiment of this application, a method for obtaining the disparity costs of the two eyes may be: inputting the hidden variables of the two eyes to at least two fully connected layers, and the two fully connected layers output the disparity costs of the two eyes. In this way, the disparity costs of the two eyes may be obtained by using fully connected layers, thereby improving feasibility and operability of the solution.

**[0062]** Optionally, based on the third embodiment corresponding to FIG. 2, in a sixth optional embodiment of the depth information determining method provided in the embodiments of this application, the calculating a t^th predicted left eye disparity value according to the t^th left eye disparity cost may include:
calculating the t^th predicted left eye disparity value in the following manner:

$$d'* = \sum_{d=1}^{D\max} d' * \sigma\left(-c'_d\right),$$

where

*d*'* represents the t^th predicted left eye disparity value, $D_{max}$ represents a maximum quantity in different disparity maps, *d*' represents a t^th left eye disparity value, $\sigma$ represents a sigmoid function, and $c'_d$ represents the t^th left eye disparity cost; and

the calculating a t^th predicted right eye disparity value according to the t^th right eye disparity cost includes: calculating the t^th predicted right eye disparity value in the following manner:

$$d''* = \sum_{d=1}^{D\max} d''* \sigma\left(-c''_{d}\right),$$

where

*d*''* represents the t^th predicted right eye disparity value, $c''_d$ represents the t^th right eye disparity cost, and *d*'' represents a t^th right eye disparity value, $c\ ''_d$ represents the t^th right eye disparity cost.

**[0063]** In this embodiment, the disparity costs of the two eyes with the size of $H*W*D_{max}$ are obtained by using a convolutional layer. A tensor form of the disparity costs of the two eyes is used, and softmax standardization is applied to tensors, so that probability tensors reflect feasible difference probabilities of all pixels. Finally, a differential argmin layer may be used to generate predicted disparity values for all differences on which probability weighting of the differential argmin layer is performed. In mathematics, the foregoing formulas describe how to obtain the predicted disparity values *d*'* (or *d*''*) of the two eyes based on given feasible disparity costs $c'_d$ (or $c''_d$) by using cost tensors of specific pixels.

**[0064]** Further, in this embodiment of this application, a specific manner of calculating predicted disparity values of two eyes is provided. That is, the predicted disparity values of the two eyes may be calculated by using a maximum quantity in different disparity maps and a left eye disparity value. In this way, a specific basis for implementing the solution is provided, thereby improving practicability and operability of the solution.

**[0065]** Optionally, based on any one of the fourth embodiment to the sixth embodiment corresponding to FIG. 2, in a seventh optional embodiment of the depth information determining method provided in the embodiments of this application, the determining first depth information according to the t^th left eye disparity map may include: calculating the first depth information in the following manner:

$$Z' = \frac{Bf}{d'*},$$

where

*Z*' represents the first depth information, *d*'* represents the t^th predicted left eye disparity value, *B* represents a binocular camera spacing, and *f* represents a focal length; and

the determining second depth information according to the t^th right eye disparity map may include: calculating the second depth information in the following manner:

$$Z'' = \frac{Bf}{d''*},$$

where

*Z*'' represents the second depth information, and *d*''* represents the t^th predicted right eye disparity value.

**[0066]** In this embodiment, after the disparity maps of the two eyes are obtained, depth information of the two eyes may be respectively calculated by using the disparity maps of the two eyes. Using calculation of first depth information of a left view as an example, a spacing of the binocular camera and a focal length of the binocular camera need to be obtained, and then, a multiplication result of the spacing and the focal length of the binocular camera is divided by an obtained predicted left eye disparity value, so that the first depth information of the left view may be obtained.

**[0067]** The following describes a deriving manner of the foregoing formula. It is assumed that internal parameters, such as a focal length and a lens, of two cameras are the same. For convenience of math description, coordinates need to be introduced. The coordinates are artificially introduced, so that an object in the real world may be located in different coordinate systems. It is assumed that X axis directions of the two cameras are the same, and image planes of the two cameras overlap. A coordinate system of a left camera is used as the basis, and a right camera is simply translated relative to the left camera, representing by coordinates ($T_x$, 0, 0). $T_x$ is generally referred to as a base line. Projection coordinates, respectively on left and right image planes, of a point P(X, Y, Z) in the space are easily obtained according to a triangular similarity relationship. Therefore, a calculation manner of a disparity can be obtained as follows:

$$d = x1 - x2 = f\frac{X}{Z} - \left(f\frac{X}{Z} - f\frac{Tx}{Z}\right),$$

and

$$d = \frac{Bf}{Z}.$$

**[0068]** The following is obtained through deriving:

$$\Rightarrow Z = \frac{Bf}{d}.$$

**[0069]** Obviously, the depth information is inversely proportional to the disparity, which is consistent with an experience result obtained by using a finger. Therefore, a near object is seen to move faster than a far object.

**[0070]** Further, in this embodiment of this application, a manner of calculating depth information is described, and depth information of two eyes may be predicted by using predicted disparity values obtained through prediction and a spacing and focal length of a binocular camera. In this way, left eye depth information and right eye depth information may be simultaneously obtained through calculation, and needed depth information may be selected according to an actual requirement, thereby improving practicability and operability of the solution.

**[0071]** The following describes the depth information determining apparatus in this application in detail. Reference is made to FIG. 7 which is a schematic diagram of an embodiment of a depth information determining apparatus equipped with a binocular camera according to an embodiment of this application. The depth information determining apparatus 20 includes:

an obtaining module 201, configured to obtain a t<sup>th</sup> left eye matching similarity from a left eye image to a right eye image, and a t<sup>th</sup> right eye matching similarity from the right eye image to the left eye image, t being an integer greater than 1;

a processing module 202, configured to process the t<sup>th</sup> left eye matching similarity obtained by the obtaining module 201 and a (t-1)<sup>th</sup> left eye attention map by using a neural network model, to obtain a t<sup>th</sup> left eye disparity map,

a processing module 202 being further configured to process the t<sup>th</sup> right eye matching similarity obtained by the obtaining module 201 and a (t-1)<sup>th</sup> right eye attention map by using the neural network model, to obtain a t<sup>th</sup> right eye disparity map; and

a determining module 203, configured to determine first depth information according to the t<sup>th</sup> left eye disparity map obtained by the processing module 202 through processing, and determine second depth information according to the t<sup>th</sup> right eye disparity map obtained by the processing module 202 through processing.

**[0072]** In this embodiment, the obtaining module 201 obtains a t<sup>th</sup> left eye matching similarity from a left eye image to a right eye image, and a t<sup>th</sup> right eye matching similarity from the right eye image to the left eye image, t being an integer greater than 1. The processing module 202 processes the t<sup>th</sup> left eye matching similarity obtained by the obtaining module 201 and a (t-1)<sup>th</sup> left eye attention map by using a neural network model, to obtain a t<sup>th</sup> left eye disparity map, and the processing module 202 processes the t<sup>th</sup> right eye matching similarity obtained by the obtaining module 201 and a (t-1)<sup>th</sup> right eye attention map by using the neural network model, to obtain a t<sup>th</sup> right eye disparity map. The determining module 203 determines first depth information according to the t<sup>th</sup> left eye disparity map obtained by the processing module 202 through processing, and determines second depth information according to the t<sup>th</sup> right eye disparity map obtained by the processing module 202 through processing.

**[0073]** In this embodiment of this application, a depth information determining apparatus is provided. Disparity maps of two eyes may be obtained by using a neural network model and attention maps of the two eyes obtained through previous learning, and then current attention maps of the two eyes obtained through learning according to the currently obtained disparity maps of the two eyes are used for next disparity maps of the two eyes. Through such recursive learning, the disparity maps of the two eyes may be continuously corrected by fully using complementary information of the two eyes, which effectively reduces an error of depth information for a region difficult to match between the two eyes.

**[0074]** Optionally, based on the embodiment corresponding to FIG. 7, referring to FIG. 8, in another embodiment of the depth information determining apparatus 20 provided in the embodiments of this application, the depth information determining apparatus 20 further includes: a mapping module 204 and a generation module 205.

**[0075]** The mapping module 204 is configured to map the t<sup>th</sup> right eye disparity map to a left eye coordinate system, to obtain a t<sup>th</sup> left eye mapping disparity map.

**[0076]** The generation module 205 is configured to generate a t<sup>th</sup> left eye attention map according to the t<sup>th</sup> left eye mapping disparity map obtained by the mapping module 204 through mapping and the t<sup>th</sup> left eye disparity map.

**[0077]** The mapping module 204 is further configured to map the t<sup>th</sup> left eye disparity map to a right eye coordinate system, to obtain a t<sup>th</sup> right eye mapping disparity map.

**[0078]** The generation module 205 is further configured to generate a t<sup>th</sup> right eye attention map according to the t<sup>th</sup> right eye mapping disparity map obtained by the mapping module 204 through mapping and the t<sup>th</sup> right eye disparity map.

**[0079]** In addition, in this embodiment of this application, the depth information determining apparatus maps the t<sup>th</sup> right eye disparity map to a left eye coordinate system, to obtain a t<sup>th</sup> left eye mapping disparity map, and generates a t<sup>th</sup> left eye attention map according to the t<sup>th</sup> left eye mapping disparity map and the t<sup>th</sup> left eye disparity map. Similarly, a t<sup>th</sup> right eye attention map may also be obtained. In this way, an attention map obtained through current recursive learning can be used for next recursive disparity prediction, and a network can purposely correct and update a disparity value of a pixel in a low-confidence region automatically detected by the current recursion according to this, thereby improving reliability of the attention maps of the two eyes.

**[0080]** Optionally, based on the embodiment corresponding to FIG. 8, in another embodiment of the depth information determining apparatus 20 provided in the embodiments of this application,

the obtaining module 201 is further configured to obtain a (t+1)<sup>th</sup> left eye matching similarity from the left eye image to the right eye image, and a (t+1)<sup>th</sup> right eye matching similarity from the right eye image to the left eye image after the determining module 203 determines the first depth information according to the t<sup>th</sup> left eye disparity map, and determines the second depth information according to the t<sup>th</sup> right eye disparity map;

the processing module 202 is further configured to process the (t+1)<sup>th</sup> left eye matching similarity and the t<sup>th</sup> left eye attention map by using the neural network model, to obtain a (t+1)<sup>th</sup> left eye disparity map;

the processing module 202 is further configured to process the (t+1)<sup>th</sup> right eye matching similarity and the t<sup>th</sup> right eye attention map by using the neural network model, to obtain a (t+1)<sup>th</sup> right eye disparity map; and

the determining module 203 is further configured to determine third depth information according to the (t+1)<sup>th</sup> left eye disparity map obtained by the processing module 202 through processing, and determine fourth depth information according to the (t+1)<sup>th</sup> right eye disparity map obtained by the processing module 202 through processing.

**[0081]** Further, in this embodiment of this application, after obtaining current depth information of the two eyes, the depth information determining apparatus may further continue to obtain next depth information of the two eyes. In this way, for comparison between the left and right eyes, a convolutional layer and an aggregation layer may be added to the neural network model, so as to generate attention maps of the two eyes. The attention maps of the two eyes are used as an input of a next step, and an LRCR model is started. Left-right mismatching regions may be focused more in the next step, thereby improving prediction accuracy.

**[0082]** Optionally, based on the embodiment corresponding to FIG. 7 or FIG. 8, in another embodiment of the depth information determining apparatus 20 provided in the embodiments of this application,

the processing module 202 is configured to obtain a t<sup>th</sup> left eye hidden variable through calculation according to the t<sup>th</sup> left eye matching similarity and the (t-1)<sup>th</sup> left eye attention map by using a ConvLSTM network;

obtain a t<sup>th</sup> left eye disparity cost according to the t<sup>th</sup> left eye hidden variable; and

calculate a t<sup>th</sup> predicted left eye disparity value according to the t<sup>th</sup> left eye disparity cost, the t<sup>th</sup> predicted left eye disparity value being used for generating the t<sup>th</sup> left eye disparity map; and

the processing module 202 is configured to obtain a t<sup>th</sup> right eye hidden variable through calculation according to the t<sup>th</sup> right eye matching similarity and the (t-1)<sup>th</sup> right eye attention map by using the ConvLSTM network;

obtain a t<sup>th</sup> right eye disparity cost according to the t<sup>th</sup> right eye hidden variable; and

calculate a t<sup>th</sup> predicted right eye disparity value according to the t<sup>th</sup> right eye disparity cost, the t<sup>th</sup> predicted right eye disparity value being used for generating the t<sup>th</sup> right eye disparity map.

**[0083]** Further, in this embodiment of this application, the ConvLSTM network is used to process the t<sup>th</sup> left eye matching similarity and the (t-1)<sup>th</sup> left eye attention map, to obtain the t<sup>th</sup> left eye disparity map, and process the t<sup>th</sup> right eye matching similarity and the (t-1)<sup>th</sup> right eye attention map, to obtain the t<sup>th</sup> right eye disparity map. In this way, based on predicted matching similarities of the two eyes, recursive prediction is performed on the disparity maps of the two eyes by using the ConvLSTM network. The ConvLSTM network has strong capabilities of sequence modeling and information processing as a conventional recursive neural network, and also can effectively extract information in neighborhood of each pixel space, so as to integrate spatial context information.

**[0084]** Optionally, based on the embodiment corresponding to FIG. 7 or FIG. 8, in another embodiment of the depth information determining apparatus 20 provided in the embodiments of this application,

the processing module 202 is configured to calculate the t^th left eye hidden variable in the following manner:

$$i'_t=\sigma(W_{xi}*X'_t+W_{hi}*H'_{t-1}+W_{ci}\circ C'_{t-1}+b_i);$$

$$f'_t=\sigma(W_{xf}*X'_t+W_{hf}*H'_{t-1}+W_{cf}\circ C'_{t-1}+b_f);$$

$$o'_t=\sigma(W_{xo}*X'_t+W_{ho}*H'_{t-1}+W_{co}\circ C'_{t-1}+b_o);$$

$$C'_t=f'_t\circ C'_{t-1}+i'_t\circ \tanh(W_{xc}*X'_t+W_{hc}*H'_{t-1}+b_c);$$

and

$$H'_t=o'_t\circ \tanh(C'_t),$$

where

$i'_t$ represents a network input gate of a t^th left eye recursion, * represents multiplication of vectors, ° represents a convolution operation, $\sigma$ represents a sigmoid function, $W_{xi}$, $W_{hi}$, $W_{ci}$, and $b_i$ represent model parameters of the network input gate, $X'_t$ represents the t^th left eye matching similarity and the (t-1)^th left eye attention map, $f'_t$ represents a forget gate of the t^th left eye recursion, $W_{xf}$, $W_{hf}$, $W_{cf}$, and $b_f$ represent model parameters of the forget gate, o't represents an output gate of the t^th left eye recursion, $W_{xo}$, $W_{ho}$, $W_{co}$, and $b_o$ represent model parameters of the output gate, $C'_t$ represents a memory cell of the t^th left eye recursion, $C'_{t-1}$ represents a memory cell of a (t-1)^th left eye recursion, tanh represents a hyperbolic tangent, $H'_{t-1}$ represents a (t-1)^th left eye hidden variable, and $H'_t$ represents the t^th left eye hidden variable; and the processing module 202 is configured to calculate the t^th right eye hidden variable in the following manner:

$$i''_t=\sigma(W_{xi}*X''_t+W_{hi}*H''_{t-1}+W_{ci}\circ C''_{t-1}+b_i);$$

$$f''_t=\sigma(W_{xf}*X''_t+W_{hf}*H''_{t-1}+W_{cf}\circ C''_{t-1}+b_f);$$

$$o''_t=\sigma(W_{xo}*X''_t+W_{ho}*H''_{t-1}+W_{co}\circ C''_{t-1}+b_o);$$

$$C''_t=f'_t\circ C''_{t-1}+i'_t\circ \tanh(W_{xc}*X''_t+W_{hc}*H''_{t-1}+b_c);$$

and

$$H''_t=o''_t\circ \tanh(C''_t),$$

where

$i''_t$ represents a network input gate of a t^th right eye recursion, $X''_t$ represents the t^th right eye matching similarity and the (t-1)^th right eye attention map, $f'_t$ represents a forget gate of the t^th right eye recursion, $o'_t$ represents an output gate of the t^th right eye recursion, $C''_t$ represents a memory cell of the t^th right eye recursion, $C''_{t-1}$ represents a memory cell of a (t-1)^th right eye recursion, $H''_{t-1}$ represents a (t-1)^th right eye hidden variable, and $H''_t$ represents the t^th right eye hidden variable.

**[0085]** Further, in this embodiment of this application, a specific manner of calculating the t^th left eye hidden variable and the t^th right eye hidden variable is described. The hidden variables of the two eyes may be obtained by using a calculation relationship provided by the ConvLSTM network. In this way, reliability of hidden variable calculation can be effectively improved, and an operability basis for implementing the solution is provided.

**[0086]** Optionally, based on the embodiment corresponding to FIG. 7 or FIG. 8, in another embodiment of the depth information determining apparatus 20 provided in the embodiments of this application,

the processing module 202 is configured to process the t^th^ left eye hidden variable by using at least two fully connected layers, to obtain the t^th^ left eye disparity cost; and

the processing module 202 is configured to process the t^th^ right eye hidden variable by using the at least two fully connected layers, to obtain the t^th^ right eye disparity cost.

**[0087]** Further, in this embodiment of this application, a method for obtaining the disparity costs of the two eyes may be: inputting hidden variables of the two eyes to at least two fully connected layers, and the two fully connected layers output the disparity costs of the two eyes. In this way, the disparity costs of the two eyes may be obtained by using fully connected layers, thereby improving feasibility and operability of the solution.

**[0088]** Optionally, based on the embodiment corresponding to FIG. 7 or FIG. 8, in another embodiment of the depth information determining apparatus 20 provided in the embodiments of this application,

the processing module 202 is configured to calculate the t^th^ predicted left eye disparity value in the following manner:

$$d'* = \sum_{d=1}^{D\max} d'* \sigma\left(-c'_d\right),$$

where

*d'** represents the t^th^ predicted left eye disparity value, $D_{max}$ represents a maximum quantity in different disparity maps, *d'* represents a t^th^ left eye disparity value, σ represents a sigmoid function, and $c'_d$ represents the t^th^ left eye disparity cost; and

the processing module 202 is configured to calculate the t^th^ predicted right eye disparity value in the following manner:

$$d"* = \sum_{d=1}^{D\max} d"* \sigma\left(-c"_d\right),$$

where

*d"** represents the t^th^ predicted right eye disparity value, $c"_d$ represents the t^th^ right eye disparity cost, and *d"* represents a t^th^ right eye disparity value, $c"_d$ represents the t^th^ right eye disparity cost.

**[0089]** Further, in this embodiment of this application, a specific manner of calculating predicted disparity values of two eyes is provided. That is, the predicted disparity values of the two eyes may be calculated by using a maximum quantity in different disparity maps and a left eye disparity value. In this way, a specific basis for implementing the solution is provided, thereby improving practicability and operability of the solution.

**[0090]** Optionally, based on the embodiment corresponding to FIG. 7 or FIG. 8, in another embodiment of the depth information determining apparatus 20 provided in the embodiments of this application,

the determining module 203 is configured to calculate the first depth information in the following manner:

$$Z' = \frac{Bf}{d'*},$$

where

Z' represents the first depth information, *d'** represents the t^th^ predicted left eye disparity value, *B* represents a binocular camera spacing, and *f* represents a focal length; and

the determining second depth information according to the t^th^right eye disparity map includes:

the determining module 203 is configured to calculate the second depth information in the following manner:

$$Z" = \frac{Bf}{d"*},$$

where

*Z"* represents the second depth information, and *d"** represents the t^th^ predicted right eye disparity value.

**[0091]** Further, in this embodiment of this application, a manner of calculating depth information is described, and depth information of two eyes may be predicted by using predicted disparity values obtained through prediction and a spacing and focal length of a binocular camera. In this way, left eye depth information and right eye depth information may be simultaneously obtained through calculation, and needed depth information may be selected according to an actual requirement, thereby improving practicability and operability of the solution.

**[0092]** FIG. 9 is a schematic structural diagram of a depth information determining apparatus according to an embodiment of this application. The depth information determining apparatus 300 may vary greatly due to different configurations or performance, and may include one or more central processing units (CPUs) 322 (for example, one or more processors), a memory 332, and one or more storage media 330 (for example, one or more mass storage devices) that store an application program 342 or data 344. The memory 332 and the storage medium 330 may be transient storages or persistent storages. The program stored in the storage medium 330 may include one or more modules (not shown in the figure), and each module may include a series of instructions and operations for the depth information determining apparatus. Still further, the CPU 322 may be configured to communicate with the storage medium 330, and perform, on the depth information determining apparatus 300, the series of instructions and operations in the storage medium 330.

**[0093]** The depth information determining apparatus 300 may further include one or more power supplies 326, one or more wired or wireless network interfaces 350, one or more input/output interfaces 358, and/or one or more operating systems 341, for example, Windows ServerTM, Mac OS XTM, UnixTM, LinuxTM, or FreeBSDTM.

**[0094]** Steps performed by the depth information determining apparatus in the foregoing embodiment may be based on the structure of the depth information determining apparatus shown in FIG. 9.

**[0095]** The CPU 322 is configured to perform the following steps:

obtaining a t$^{th}$ left eye matching similarity from a left eye image to a right eye image, and a t$^{th}$ right eye matching similarity from the right eye image to the left eye image, t being an integer greater than 1;

processing the t$^{th}$ left eye matching similarity and a (t-1)$^{th}$ left eye attention map by using a neural network model, to obtain a t$^{th}$ left eye disparity map;

processing the t$^{th}$ right eye matching similarity and a (t-1)$^{th}$ right eye attention map by using the neural network model, to obtain a t$^{th}$ right eye disparity map; and

determining first depth information according to the t$^{th}$ left eye disparity map, and determining second depth information according to the t$^{th}$ right eye disparity map.

**[0096]** Optionally, the CPU 322 is further configured to perform the following steps:

mapping the t$^{th}$ right eye disparity map to a left eye coordinate system, to obtain a t$^{th}$ left eye mapping disparity map;

generating a t$^{th}$ left eye attention map according to the t$^{th}$ left eye mapping disparity map and the t$^{th}$ left eye disparity map;

mapping the t$^{th}$ left eye disparity map to a right eye coordinate system, to obtain a t$^{th}$ right eye mapping disparity map; and

generating a t$^{th}$ right eye attention map according to the t$^{th}$ right eye mapping disparity map and the t$^{th}$ right eye disparity map.

**[0097]** Optionally, the CPU 322 is further configured to perform the following steps:

obtaining a (t+1)$^{th}$ left eye matching similarity from the left eye image to the right eye image, and a (t+1)$^{th}$ right eye matching similarity from the right eye image to the left eye image;
processing the (t+1)$^{th}$ left eye matching similarity and the t$^{th}$ left eye attention map by using the neural network model, to obtain a (t+1)$^{th}$ left eye disparity map;
processing the (t+1)$^{th}$ right eye matching similarity and the t$^{th}$ right eye attention map by using the neural network model, to obtain a (t+1)$^{th}$ right eye disparity map; and
determining third depth information according to the (t+1)$^{th}$ left eye disparity map, and determining fourth depth information according to the (t+1)$^{th}$ right eye disparity map.

**[0098]** Optionally, the CPU 322 is configured to perform the following steps:

obtaining a t$^{th}$ left eye hidden variable through calculation according to the t$^{th}$ left eye matching similarity and the

(t-1)th left eye attention map by using a ConvLSTM network;
obtaining a tth left eye disparity cost according to the tth left eye hidden variable; and
calculating a tth predicted left eye disparity value according to the tth left eye disparity cost, the tth predicted left eye disparity value being used for generating the tth left eye disparity map; and
obtaining a tth right eye hidden variable through calculation according to the tth right eye matching similarity and the (t-1)th right eye attention map by using the ConvLSTM network;
obtaining a tth right eye disparity cost according to the tth right eye hidden variable; and
calculating a tth predicted right eye disparity value according to the tth right eye disparity cost, the tth predicted right eye disparity value being used for generating the tth right eye disparity map.

**[0099]** Optionally, the CPU 322 is configured to perform the following steps:
calculating the tth left eye hidden variable in the following manner:

$$i'_t=\sigma(W_{xi}*X'_t+W_{hi}*H'_{t-1}+W_{ci}{}^\circ C'_{t-1}+b_i);$$

$$f'_t=\sigma(W_{xf}*X'_t+W_{hf}*H'_{t-1}+W_{cf}{}^\circ C'_{t-1}+b_f);$$

$$o'_t=\sigma(W_{xo}*X'_t+W_{ho}*H'_{t-1}+W_{co}{}^\circ C'_{t-1}+b_o);$$

$$C'_t=f'_t{}^\circ C'_{t-1}+i'_t{}^\circ \tanh(W_{xc}*X'_t+W_{hc}*H'_{t-1}+b_c);$$

and

$$H'_t=o'_t{}^\circ \tanh(C'_t),$$

where
$i'_t$ represents a network input gate of a tth left eye recursion, * represents multiplication of vectors, ° represents a convolution operation, $\sigma$ represents a sigmoid function, $W_{xi}$, $W_{hi}$, $W_{ci}$, and $b_i$ represent model parameters of the network input gate, $X'_t$ represents the tth left eye matching similarity and the (t-1)th left eye attention map, $f'_t$ represents a forget gate of the tth left eye recursion, $W_{xf}$, $W_{hf}$, $W_{cf}$, and $b_f$ represent model parameters of the forget gate, o't represents an output gate of the tth left eye recursion, $W_{xo}$, $W_{ho}$, $W_{co}$, and $b_o$ represent model parameters of the output gate, C't represents a memory cell of the tth left eye recursion, $C'_{t-1}$ represents a memory cell of a (t-1)th left eye recursion, tanh represents a hyperbolic tangent, $H'_{t-1}$ represents a (t-1)th left eye hidden variable, and $H'_t$ represents the tth left eye hidden variable; and
calculating the tth right eye hidden variable in the following manner:

$$i''_t=\sigma(W_{xi}*X''_t+W_{hi}*H''_{t-1}+W_{ci}{}^\circ C''_{t-1}+b_i);$$

$$f''_t=\sigma(W_{xf}*X''_t+W_{hf}*H''_{t-1}+W_{cf}{}^\circ C''_{t-1}+b_f);$$

$$o''_t=\sigma(W_{xo}*X''_t+W_{ho}*H''_{t-1}+W_{co}{}^\circ C''_{t-1}+b_o);$$

$$C''_t=f'_t{}^\circ C''_{t-1}+i'_t{}^\circ \tanh(W_{xc}*X''_t+W_{hc}*H''_{t-1}+b_c);$$

and

$$H''_t=o''_t{}^\circ \tanh(C''_t),$$

where
i"$_t$ represents a network input gate of a t$^{th}$ right eye recursion, X"$_t$ represents the t$^{th}$ right eye matching similarity and the (t-1)$^{th}$ right eye attention map, f'$_t$ represents a forget gate of the t$^{th}$ right eye recursion, o'$_t$ represents an output gate of the t$^{th}$ right eye recursion, C"$_t$ represents a memory cell of the t$^{th}$ right eye recursion, C"$_{t-1}$ represents a memory cell of a (t-1)$^{th}$ right eye recursion, H"$_{t-1}$ represents a (t-1)$^{th}$ right eye hidden variable, and H"$_t$ represents the t$^{th}$ right eye hidden variable.

**[0100]** Optionally, the CPU 322 is configured to perform the following steps:

processing the t$^{th}$ left eye hidden variable by using at least two fully connected layers, to obtain the t$^{th}$ left eye disparity cost; and
processing the t$^{th}$ right eye hidden variable by using the at least two fully connected layers, to obtain the t$^{th}$ right eye disparity cost.

**[0101]** Optionally, the CPU 322 is configured to perform the following steps:
calculating the t$^{th}$ predicted left eye disparity value in the following manner:

$$d'^{*}=\sum\nolimits_{d=1}^{D\max} d'^{*}\sigma\left(-c'_{d}\right),$$

where
$d'^{*}$ represents the t$^{th}$ predicted left eye disparity value, $D_{max}$ represents a maximum quantity in different disparity maps, $d'$ represents a t$^{th}$ left eye disparity value, $\sigma$ represents a sigmoid function, and $c'_d$ represents the t$^{th}$ left eye disparity cost; and
calculating the t$^{th}$ predicted right eye disparity value in the following manner:

$$d''^{*}=\sum\nolimits_{d=1}^{D\max} d''^{*}\sigma\left(-c''_{d}\right),$$

where
$d''^{*}$ represents the t$^{th}$ predicted right eye disparity value, $c''_d$ represents the t$^{th}$ right eye disparity cost, and $d''$ represents a t$^{th}$ right eye disparity value, $c''_d$ represents the t$^{th}$ right eye disparity cost.

**[0102]** Optionally, the CPU 322 is configured to perform the following steps:
calculating the first depth information in the following manner:

$$Z'=\frac{Bf}{d'^{*}},$$

where
$Z'$ represents the first depth information, $d'^{*}$ represents the t$^{th}$ predicted left eye disparity value, $B$ represents a binocular camera spacing, and $f$ represents a focal length; and
calculating the second depth information in the following manner:

$$Z''=\frac{Bf}{d''^{*}},$$

where
$Z''$ represents the second depth information, and $d''^{*}$ represents the t$^{th}$ predicted right eye disparity value.

**[0103]** A person skilled in the art may clearly understand that, for the purpose of convenient and brief description, for specific work processes of the foregoing described system, apparatus, and unit, reference may be made to corresponding processes in the foregoing method embodiments, and details are not described herein again.

**[0104]** In the embodiments provided in this application, it is to be understood that the disclosed system, apparatus, and method may be implemented in other manners. For example, the described apparatus embodiment is merely an example. For example, the unit division is merely logical function division and may be other division during actual implementation. For example, a plurality of units or components may be combined or integrated into another system, or some features may be ignored or not performed. In addition, the displayed or discussed mutual couplings or direct couplings or communication connections may be implemented by using some interfaces. The indirect couplings or communication connections between the apparatuses or units may be implemented in electric, mechanical, or other forms.

**[0105]** The units described as separate parts may or may not be physically separate, and parts displayed as units may or may not be physical units, may be located in one position, or may be distributed on a plurality of network units. Some or all of the units may be selected according to actual requirements to achieve the objectives of the solutions of the embodiments.

**[0106]** In addition, functional units in the embodiments of this application may be integrated into one processing unit, or each of the units may exist alone physically, or two or more units are integrated into one unit. The integrated unit may be implemented in the form of hardware, or may be implemented in the form of software functional unit.

**[0107]** When the integrated unit is implemented in the form of a software functional unit and sold or used as an independent product, the integrated unit may be stored in a computer-readable storage medium. Based on such an understanding, the technical solutions of this application essentially, or the part contributing to the prior art, or some or all of the technical solutions may be implemented in a form of a software product. The computer software product is stored in a storage medium and includes several instructions for instructing a computer device (which may be a personal computer, a server, a network device, or the like) to perform all or some steps of the methods described in the embodiments of this application. The foregoing storage medium includes: any medium that can store program code, such as a USB flash drive, a removable hard disk, a read-only memory (ROM), a random access memory (RAM), a magnetic disk, or a compact disc.

**[0108]** The foregoing embodiments are merely intended for describing the technical solutions of this application, but not for limiting this application. Although this application is described in detail with reference to the foregoing embodiments, a person of ordinary skill in the art needs to understand that they may still make modifications to the technical solutions described in the foregoing embodiments or make equivalent replacements to some technical features thereof, without departing from the spirit and scope of the technical solutions of the embodiments of this application.

## Claims

**1.** A depth information determining method, applied to a facility equipped with a binocular camera, the method comprising:

obtaining a $t^{th}$ left eye matching similarity from a left eye image to a right eye image, and a $t^{th}$ right eye matching similarity from the right eye image to the left eye image, t being an integer greater than 1;
processing the $t^{th}$ left eye matching similarity and a $(t-1)^{th}$ left eye attention map by using a neural network model, to obtain a $t^{th}$ left eye disparity map;
processing the $t^{th}$ right eye matching similarity and a $(t-1)^{th}$ right eye attention map by using the neural network model, to obtain a $t^{th}$ right eye disparity map; and
determining first depth information according to the $t^{th}$ left eye disparity map, and determining second depth information according to the $t^{th}$ right eye disparity map.

**2.** The method according to claim 1, further comprising:

mapping the $t^{th}$ right eye disparity map to a left eye coordinate system, to obtain a $t^{th}$ left eye mapping disparity map;
generating a $t^{th}$ left eye attention map according to the $t^{th}$ left eye mapping disparity map and the $t^{th}$ left eye disparity map;
mapping the $t^{th}$ left eye disparity map to a right eye coordinate system, to obtain a $t^{th}$ right eye mapping disparity map; and
generating a $t^{th}$ right eye attention map according to the $t^{th}$ right eye mapping disparity map and the $t^{th}$ right eye disparity map.

**3.** The method according to claim 2, wherein after the determining first depth information according to the $t^{th}$ left eye disparity map, and determining second depth information according to the $t^{th}$ right eye disparity map, the method further comprises:

obtaining a $(t+1)^{th}$ left eye matching similarity from the left eye image to the right eye image, and a $(t+1)^{th}$ right eye matching similarity from the right eye image to the left eye image;
processing the $(t+1)^{th}$ left eye matching similarity and the $t^{th}$ left eye attention map by using the neural network

model, to obtain a (t+1)th left eye disparity map;
processing the (t+1)th right eye matching similarity and the tth right eye attention map by using the neural network model, to obtain a (t+1)th right eye disparity map; and
determining third depth information according to the (t+1)th left eye disparity map, and determining fourth depth information according to the (t+1)th right eye disparity map.

4. The method according to any one of claims 1 to 3, wherein the processing the tth left eye matching similarity and a (t-1)th left eye attention map by using a neural network model, to obtain a tth left eye disparity map comprises:

obtaining a tth left eye hidden variable through calculation according to the tth left eye matching similarity and the (t-1)th left eye attention map by using a convolutional long short-term memory (ConvLSTM) network;
obtaining a tth left eye disparity cost according to the tth left eye hidden variable; and
calculating a tth predicted left eye disparity value according to the tth left eye disparity cost, the tth predicted left eye disparity value being used for generating the tth left eye disparity map; and
the processing the tth right eye matching similarity and a (t-1)th right eye attention map by using the neural network model, to obtain a tth right eye disparity map comprises:

obtaining a tth right eye hidden variable through calculation according to the tth right eye matching similarity and the (t-1)th right eye attention map by using the ConvLSTM network;
obtaining a tth right eye disparity cost according to the tth right eye hidden variable; and
calculating a tth predicted right eye disparity value according to the tth right eye disparity cost, the tth predicted right eye disparity value being used for generating the tth right eye disparity map.

5. The method according to claim 4, wherein the obtaining a tth left eye hidden variable through calculation according to the tth left eye matching similarity and the (t-1)th left eye attention map by using a ConvLSTM network comprises:
calculating the tth left eye hidden variable in the following manner:

$$i'_t=\sigma(W_{xi}*X'_t+W_{hi}*H'_{t-1}+W_{ci}\circ C'_{t-1}+b_i);$$

$$f'_t=\sigma(W_{xf}*X'_t+W_{hf}*H'_{t-1}+W_{cf}\circ C'_{t-1}+b_f);$$

$$o'_t=\sigma(W_{xo}*X'_t+W_{ho}*H'_{t-1}+W_{co}\circ C'_{t-1}+b_o);$$

$$C'_t=f'_t\circ C'_{t-1}+i'_t\circ \tanh(W_{xc}*X'_t+W_{hc}*H'_{t-1}+b_c);$$

and

$$H'_t=o'_t\circ \tanh(C'_t),$$

wherein
$i'_t$ represents a network input gate of a tth left eye recursion, * represents multiplication of vectors, ° represents a convolution operation, σ represents a sigmoid function, $W_{xi}$, $W_{hi}$, $W_{ci}$, and $b_i$ represent model parameters of the network input gate, $X'_t$ represents the tth left eye matching similarity and the (t-1)th left eye attention map, $f'_t$ represents a forget gate of the tth left eye recursion, $W_{xf}$, $W_{hf}$, $W_{cf}$, and $b_f$ represent model parameters of the forget gate, o't represents an output gate of the tth left eye recursion, $W_{xo}$, $W_{ho}$, $W_{co}$, and $b_o$ represent model parameters of the output gate, C't represents a memory cell of the tth left eye recursion, $C'_{t-1}$ represents a memory cell of a (t-1)th left eye recursion, tanh represents a hyperbolic tangent, $H'_{t-1}$ represents a (t-1)th left eye hidden variable, and $H'_t$ represents the tth left eye hidden variable; and
the obtaining a tth right eye hidden variable through calculation according to the tth right eye matching similarity and the (t-1)th right eye attention map by using the ConvLSTM network comprises:
calculating the tth right eye hidden variable in the following manner:

$$i''_t=\sigma(W_{xi}*X''_t+W_{hi}*H''_{t-1}+W_{ci}\circ C''_{t-1}+b_i);$$

$$f''_t=\sigma(W_{xf}*X''_t+W_{hf}*H''_{t-1}+W_{cf}\circ C''_{t-1}+b_f);$$

$$o''_t=\sigma(W_{xo}*X''_t+W_{ho}*H''_{t-1}+W_{co}\circ C''_{t-1}+b_o);$$

$$C''_t=f'_t\circ C''_{t-1}+i'_t\circ \tanh(W_{xc}*X''_t+W_{hc}*H''_{t-1}+b_c);$$

and

$$H''_t=o''_t\circ \tanh(C''_t),$$

wherein
$i''_t$ represents a network input gate of a t<sup>th</sup> right eye recursion, $X''_t$ represents the t<sup>th</sup> right eye matching similarity and the (t-1)<sup>th</sup> right eye attention map, $f'_t$ represents a forget gate of the t<sup>th</sup> right eye recursion, $o'_t$ represents an output gate of the t<sup>th</sup> right eye recursion, $C''_t$ represents a memory cell of the t<sup>th</sup> right eye recursion, $C''_{t-1}$ represents a memory cell of a (t-1)<sup>th</sup> right eye recursion, $H''_{t-1}$ represents a (t-1)<sup>th</sup> right eye hidden variable, and $H''_t$ represents the t<sup>th</sup> right eye hidden variable.

**6.** The method according to claim 4, wherein the obtaining a t<sup>th</sup> left eye disparity cost according to the t<sup>th</sup> left eye hidden variable comprises:

processing the t<sup>th</sup> left eye hidden variable by using at least two fully connected layers, to obtain the t<sup>th</sup> left eye disparity cost; and
the obtaining a t<sup>th</sup> right eye disparity cost according to the t<sup>th</sup> right eye hidden variable comprises:
processing the t<sup>th</sup> right eye hidden variable by using the at least two fully connected layers, to obtain the t<sup>th</sup> right eye disparity cost.

**7.** The method according to claim 4, wherein the calculating a t<sup>th</sup> predicted left eye disparity value according to the t<sup>th</sup> left eye disparity cost comprises:
calculating the t<sup>th</sup> predicted left eye disparity value in the following manner:

$$d'* = \sum_{d=1}^{D\max} d'*\sigma\left(-c'_d\right),$$

wherein
$d'*$ represents the t<sup>th</sup> predicted left eye disparity value, $D_{max}$ represents a maximum quantity in different disparity maps, $d'$ represents a t<sup>th</sup> left eye disparity value, σ represents a sigmoid function, and $c'_d$ represents the t<sup>th</sup> left eye disparity cost; and
the calculating a t<sup>th</sup> predicted right eye disparity value according to the t<sup>th</sup> right eye disparity cost comprises:
calculating the t<sup>th</sup> predicted right eye disparity value in the following manner:

$$d''* = \sum_{d=1}^{D\max} d''*\sigma\left(-c''_d\right),$$

wherein
$d''*$ represents the t<sup>th</sup> predicted right eye disparity value, $c''_d$ represents the t<sup>th</sup> right eye disparity cost, $d''$ represents a t<sup>th</sup> right eye disparity value, and $c\ ''_d$ represents the t<sup>th</sup> right eye disparity cost.

**8.** The method according to any one of claims 5 to 7, wherein the determining first depth information according to the t<sup>th</sup> left eye disparity map comprises:
calculating the first depth information in the following manner:

$$Z' = \frac{Bf}{d'*},$$

wherein

Z’ represents the first depth information, $d'*$ represents the tth predicted left eye disparity value, $B$ represents a binocular camera spacing, and $f$ represents a focal length; and
the determining second depth information according to the tth right eye disparity map comprises:
calculating the second depth information in the following manner:

$$Z'' = \frac{Bf}{d''*},$$

wherein
Z" represents the second depth information, and $d''*$ represents the tth predicted right eye disparity value.

**9.** A depth information determining apparatus equipped with a binocular camera, comprising a memory, a processor, and a bus system,
the memory being configured to store a program;
the processor being configured to execute the program in the memory, to perform the following operations:

obtaining a tth left eye matching similarity from a left eye image to a right eye image, and a tth right eye matching similarity from the right eye image to the left eye image, t being an integer greater than 1;
processing the tth left eye matching similarity and a (t-1)th left eye attention map by using a neural network model, to obtain a tth left eye disparity map;
processing the tth right eye matching similarity and a (t-1)th right eye attention map by using the neural network model, to obtain a tth right eye disparity map; and
determining first depth information according to the tth left eye disparity map, and determining second depth information according to the tth right eye disparity map; and
the bus system being configured to connect the memory and the processor for communication between the memory and the processor.

**10.** The depth information determining apparatus according to claim 9, wherein
the processor is further configured to map the tth right eye disparity map to a left eye coordinate system, to obtain a tth left eye mapping disparity map;
generate a tth left eye attention map according to the tth left eye mapping disparity map and the tth left eye disparity map;
map the tth left eye disparity map to a right eye coordinate system, to obtain a tth right eye mapping disparity map; and
generate a tth right eye attention map according to the tth right eye mapping disparity map and the tth right eye disparity map.

**11.** The depth information determining apparatus according to claim 9, wherein
the processor is further configured to obtain a (t+1)th left eye matching similarity from the left eye image to the right eye image, and a (t+1)th right eye matching similarity from the right eye image to the left eye image;
process the (t+1)th left eye matching similarity and the tth left eye attention map by using the neural network model, to obtain a (t+1)th left eye disparity map;
process the (t+1)th right eye matching similarity and the tth right eye attention map by using the neural network model, to obtain a (t+1)th right eye disparity map; and
determine third depth information according to the (t+1)th left eye disparity map, and determine fourth depth information according to the (t+1)th right eye disparity map.

**12.** A computer-readable storage medium, comprising instructions, the instructions, when running on a computer, causing the computer to perform the method according to any one of claims 1 to 8.

**13.** A computer program product comprising instructions, the instructions, when running on a computer, causing the computer to perform the method according to any one of claims 1 to 8.

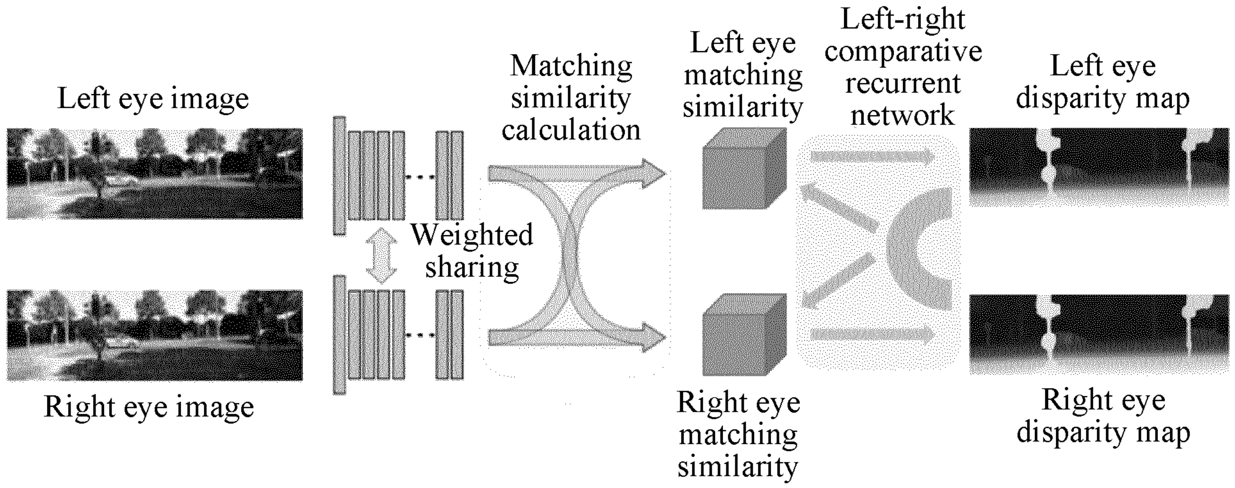
Left eye image
Right eye image
Weighted sharing
Matching similarity calculation
Left eye matching similarity
Right eye matching similarity
Left-right comparative recurrent network
Left eye disparity map
Right eye disparity map

FIG. 1A

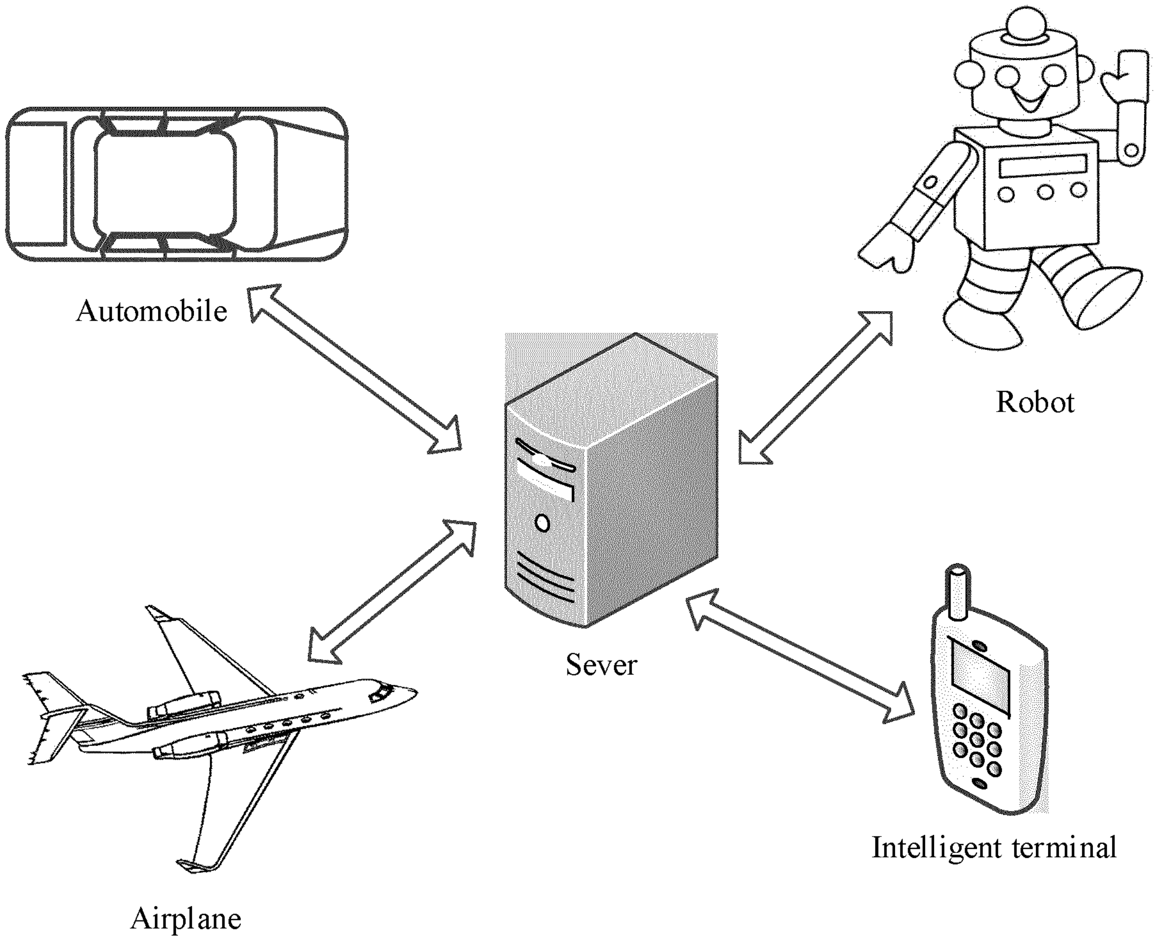
Automobile
Robot
Sever
Airplane
Intelligent terminal

FIG. 1B

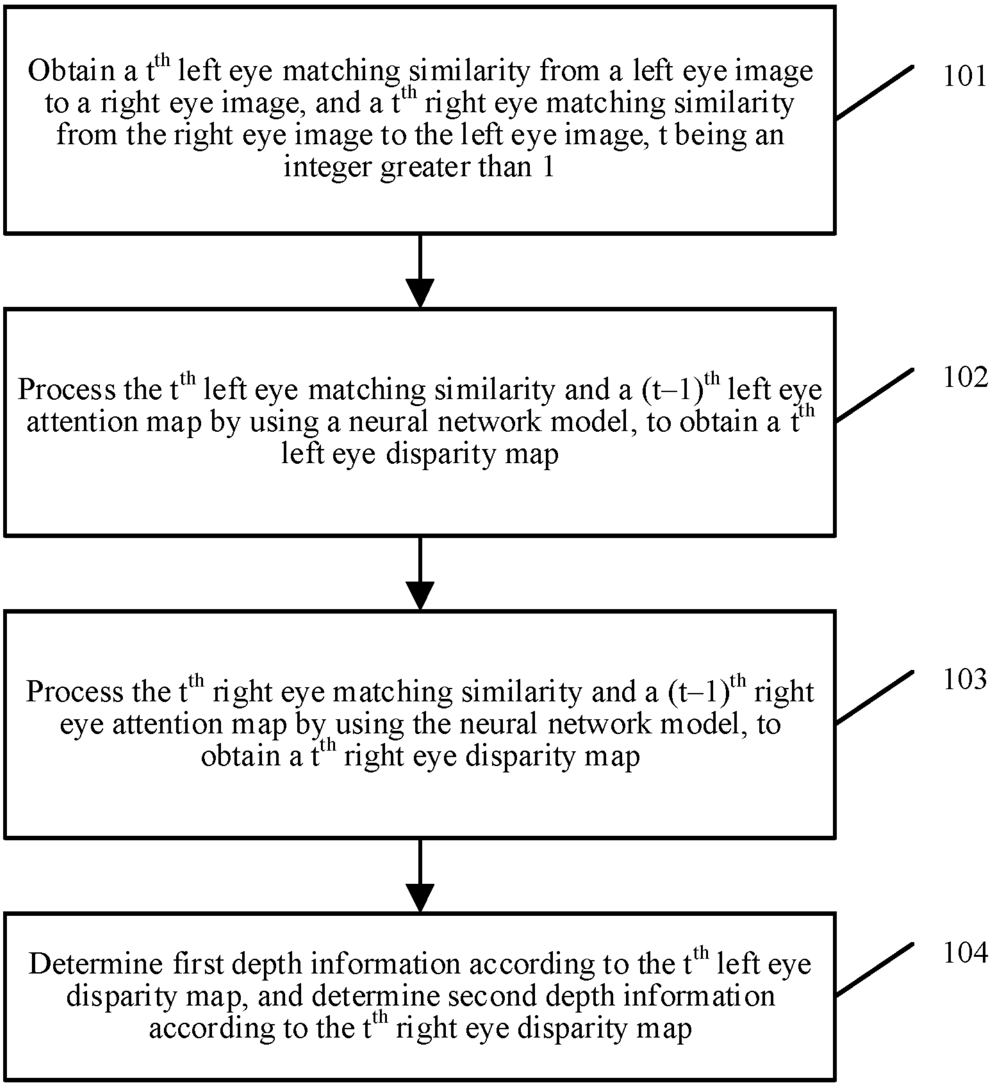
Obtain a $t^{th}$ left eye matching similarity from a left eye image to a right eye image, and a $t^{th}$ right eye matching similarity from the right eye image to the left eye image, t being an integer greater than 1
101
Process the $t^{th}$ left eye matching similarity and a $(t–1)^{th}$ left eye attention map by using a neural network model, to obtain a $t^{th}$ left eye disparity map
102
Process the $t^{th}$ right eye matching similarity and a $(t–1)^{th}$ right eye attention map by using the neural network model, to obtain a $t^{th}$ right eye disparity map
103
Determine first depth information according to the $t^{th}$ left eye disparity map, and determine second depth information according to the $t^{th}$ right eye disparity map
104

FIG. 2

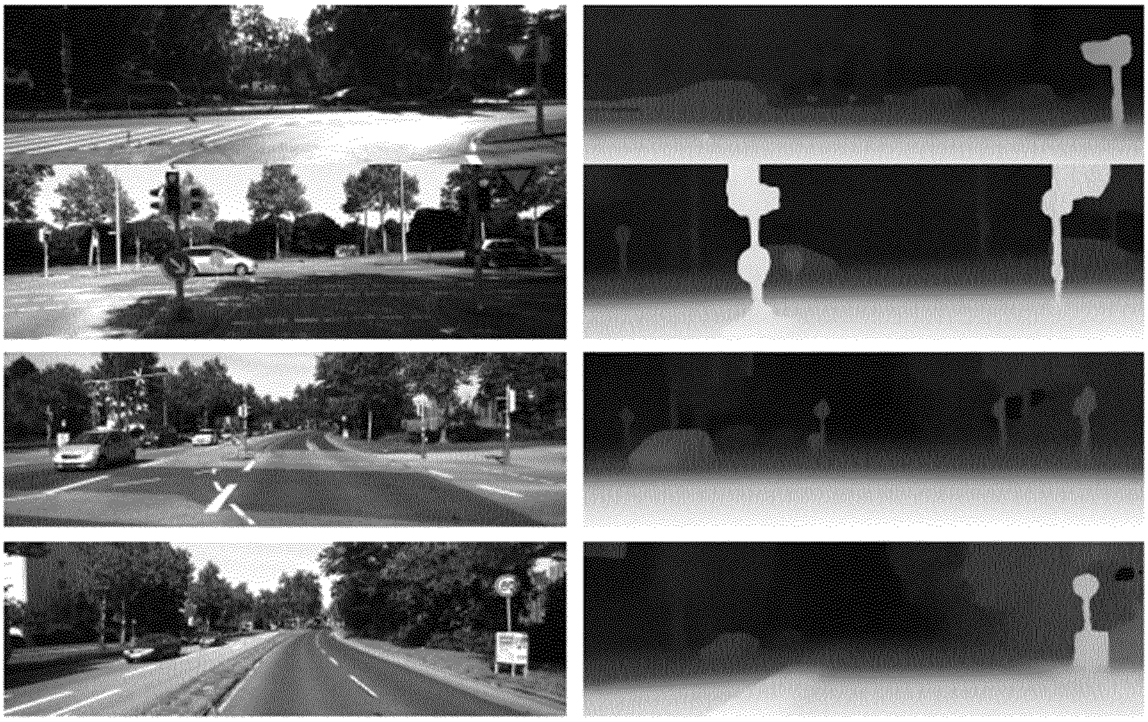

FIG. 3

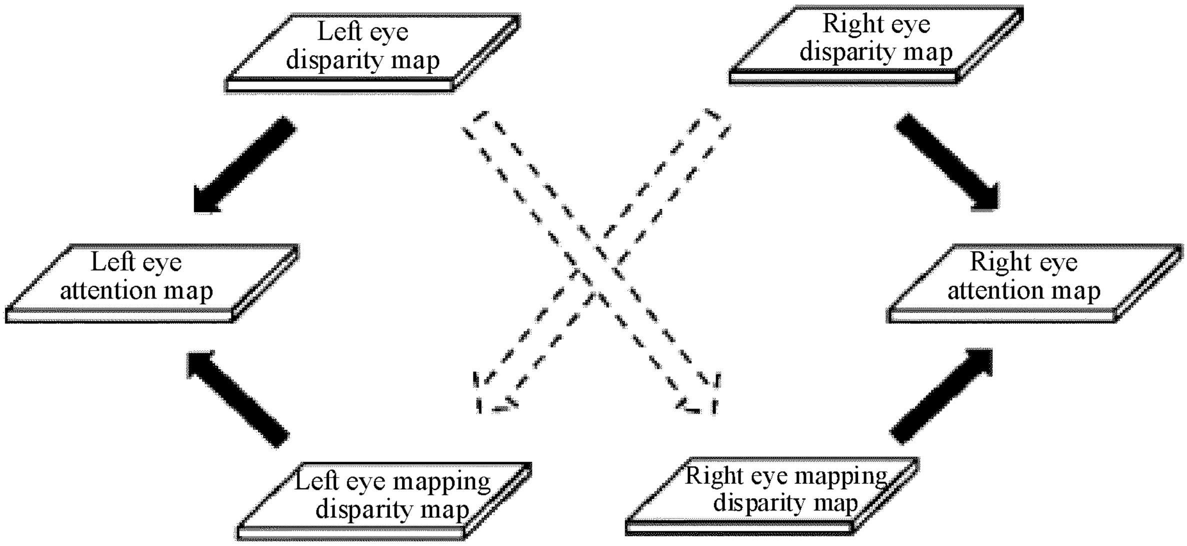
Left eye
disparity map
Right eye
disparity map
Left eye
attention map
Right eye
attention map
Left eye mapping
disparity map
Right eye mapping
disparity map

FIG. 4

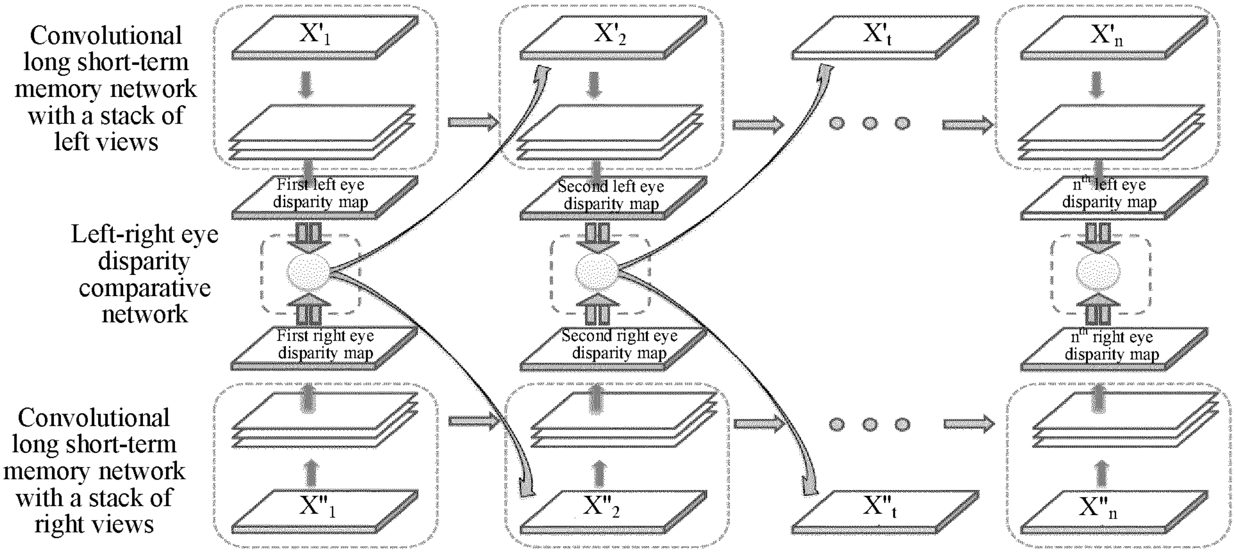
Convolutional long short-term memory network with a stack of left views
X'1
X'2
X't
X'n
First left eye disparity map
Second left eye disparity map
nth left eye disparity map
Left-right eye disparity comparative network
First right eye disparity map
Second right eye disparity map
nth right eye disparity map
Convolutional long short-term memory network with a stack of right views
X"1
X"2
X"t
X"n

FIG. 5

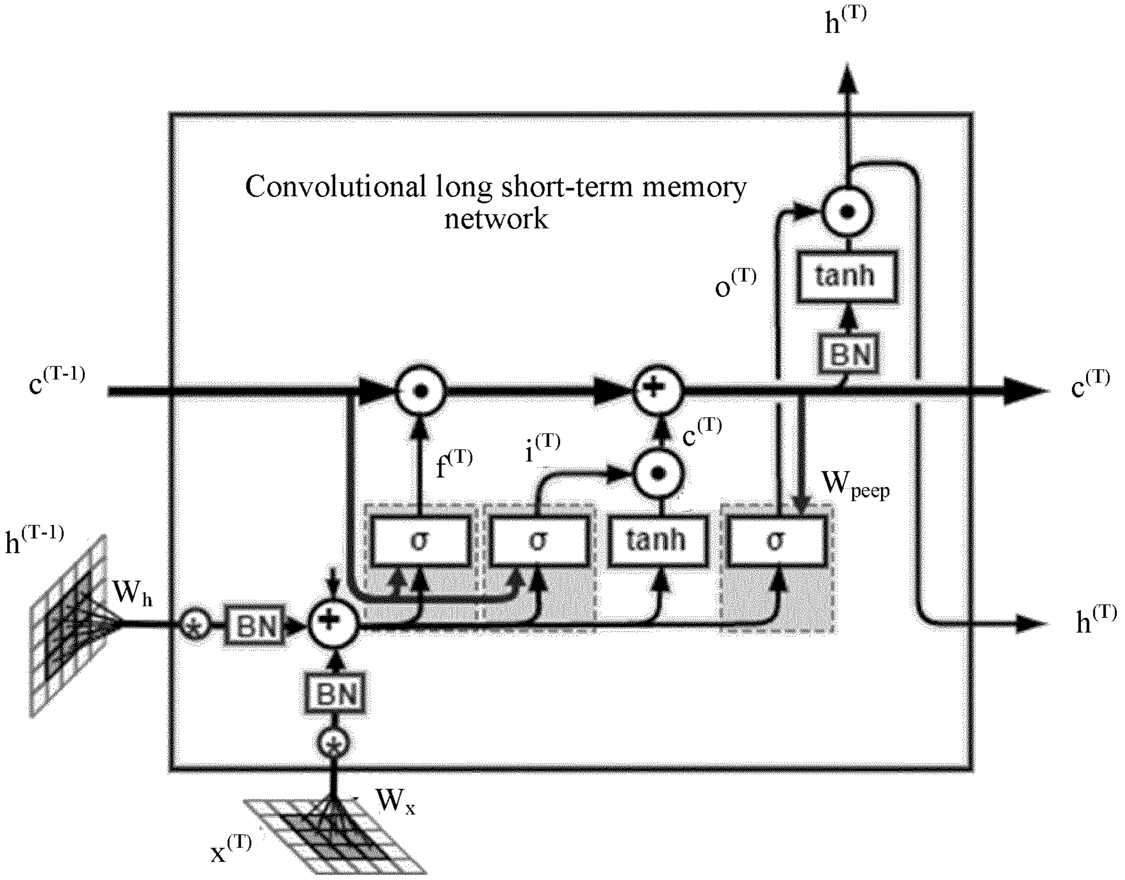
h(T)
Convolutional long short-term memory network
o(T)
tanh
BN
c(T-1)
c(T)
f(T)
i(T)
c(T)
W peep
h(T-1)
Wh
BN
σ
σ
tanh
σ
h(T)
BN
Wx
x(T)

FIG. 6

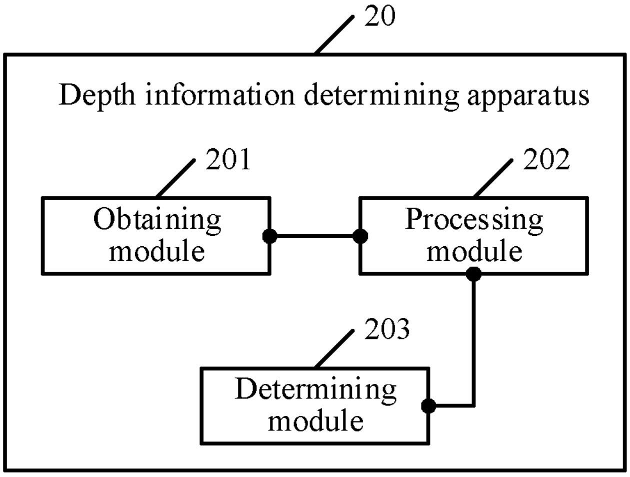
20
Depth information determining apparatus
201
Obtaining module
202
Processing module
203
Determining module

FIG. 7

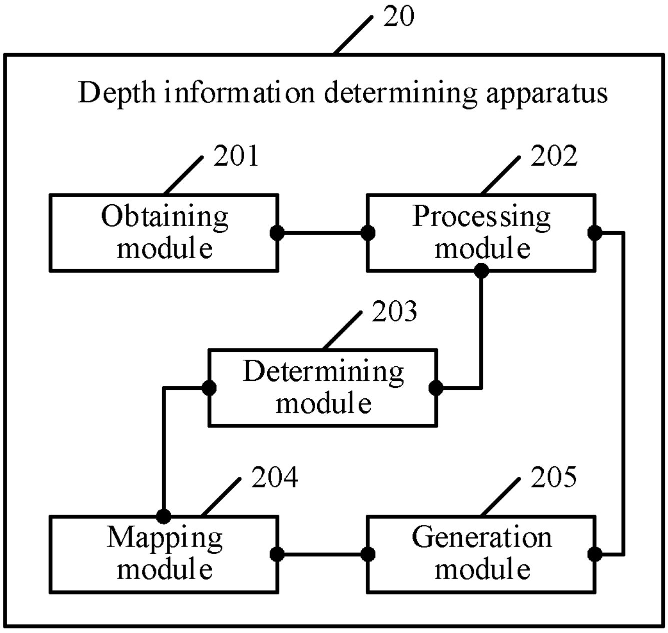
20
Depth information determining apparatus
201
Obtaining module
202
Processing module
203
Determining module
204
Mapping module
205
Generation module

FIG. 8

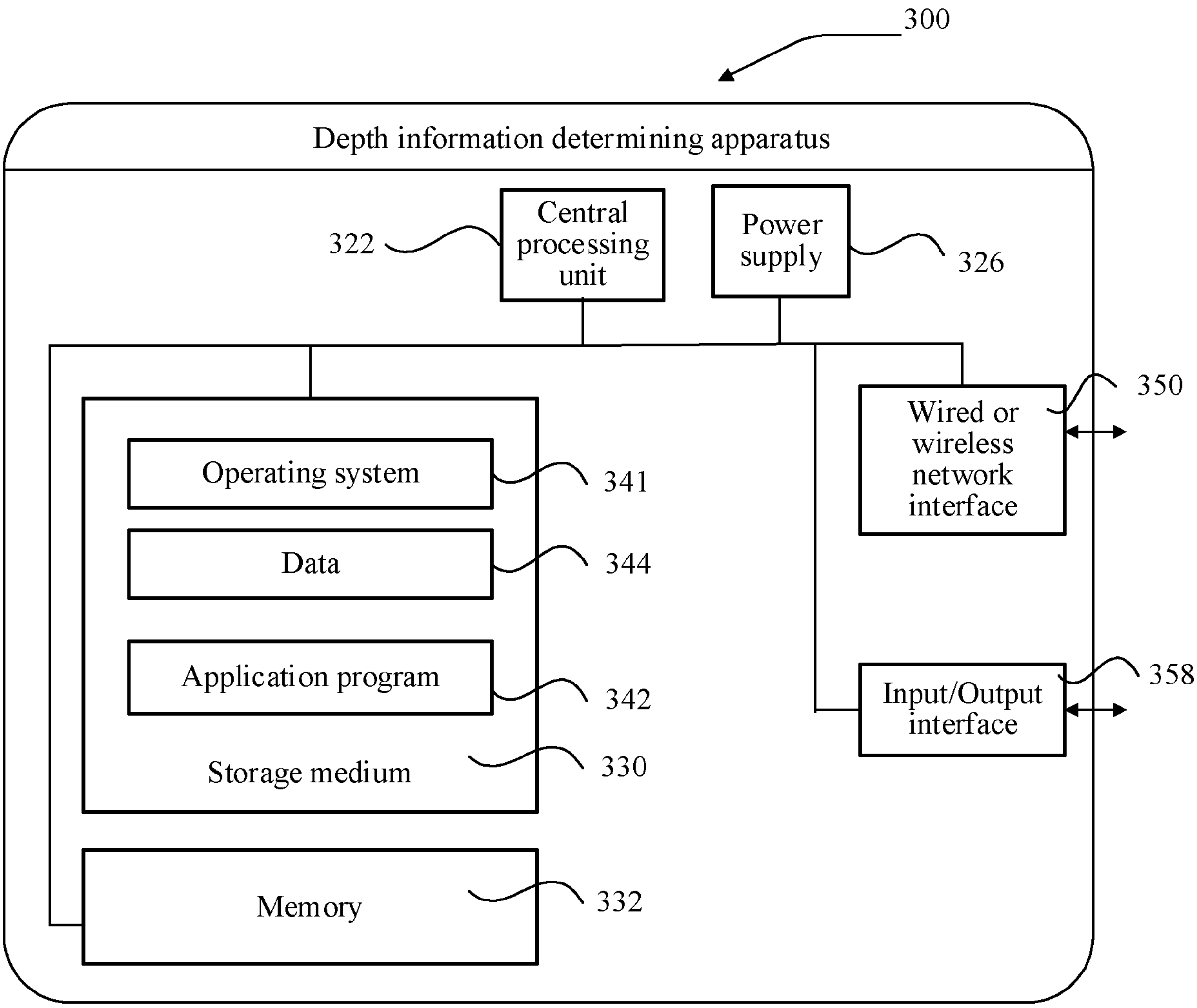
300
Depth information determining apparatus
Central processing unit
322
Power supply
326
Wired or wireless network interface
350
Operating system
341
Data
344
Application program
342
Storage medium
330
Input/Output interface
358
Memory
332

FIG. 9

| INTERNATIONAL SEARCH REPORT | International application No.<br>**PCT/CN2019/077669** |
|---|---|

**A. CLASSIFICATION OF SUBJECT MATTER**

G06T 7/55(2017.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)

G06T7/-

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS, CNTXT, CNKI, VEN, USTXT, WOTXT, EPTXT: 深度, 视差, 双目, 匹配, 相似, 神经网络, 左, 右, deep, depth, disparity, parallax, parallex, binocular, match, similarity, neural network, left, righ,

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| PX | CN 108537837 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 14 September 2018 (2018-09-14)<br>entire document | 1-13 |
| A | CN 106600650 A (HANGZHOU LANXIN TECHNOLOGY CO., LTD.) 26 April 2017 (2017-04-26)<br>description, paragraphs [0004]-[0014] | 1-13 |
| A | CN 107590831 A (UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA) 16 January 2018 (2018-01-16)<br>description, paragraphs [0021]-[0025] | 1-13 |
| A | US 2013266211 A1 (BRIGHAM YOUNG UNIVERSITY) 10 October 2013 (2013-10-10)<br>entire document | 1-13 |

□ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

| | |
|---|---|
| Date of the actual completion of the international search<br>**22 May 2019** | Date of mailing of the international search report<br>**14 June 2019** |
| Name and mailing address of the ISA/CN<br>**National Intellectual Property Administration, PRC (ISA/CN)**<br>**No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088**<br>**China** | Authorized officer |
| Facsimile No. **(86-10)62019451** | Telephone No. |

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2019/077669

| Patent document cited in search report | Publication date (day/month/year) | Patent family member(s) | Publication date (day/month/year) |
|---|---|---|---|
| CN 108537837 A | 14 September 2018 | None | |
| CN 106600650 A | 26 April 2017 | None | |
| CN 107590831 A | 16 January 2018 | None | |
| US 2013266211 A1 | 10 October 2013 | US 9317923 B2 | 19 April 2016 |

## REFERENCES CITED IN THE DESCRIPTION

*This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.*

### Patent documents cited in the description

- CN 201810301988 **[0001]**